TRANSACTIONS
OF THE ASME
Series D

Journal of BASIC ENGINEERING

June 1960

Divisions: LUBRICATION • HYDRAULIC • INSTRUMENTS AND REGULATORS • METALS ENGINEERING • **Research Committees:** AUTOMATIC REGULATION THEORY • EFFECT OF TEMPERATURE ON THE PROPERTIES OF METALS • FLUID METERS • PREVENTION OF FRACTURE IN METALS • MECHANICAL PRESSURE ELEMENTS • LUBRICATION • PLASTIC FLOW OF METALS

Published Quarterly by
THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS

TRANSACTIONS OF THE ASME
Series D

Journal of
BASIC ENGINEERING

VOLUME 82 • SERIES D • NUMBER 2
June 1960

CONTENTS

251 **The Fluids Engineering Laboratory at the University of Michigan**
G. V. Edmonson

257 **Flow Parameters in Hydrostatic Lubrication for Several Bearing Shapes**
S. Raynor and A. Charnes

265 **A General Method for Correlating Labyrinth-Seal Leak-Rate Data**
F. E. Heffner

276 **Hydrostatic Gas Bearings**
J. H. Laub

287 **Effect of Hardness, Surface Finish, and Grain Size on Rolling-Contact Fatigue Life of M-50 Bearing Steel**
R. A. Baughman

295 **The Importance of Spinning Friction in Thrust-Carrying Ball Bearings**
G. S. Reichenbach

302 **The Effect of Ball Bearing Steel Structure on Rolling Friction and Contact Plastic Deformation**
R. C. Drutowski and E. B. Mikus

309 **A General Theory for Elastically Constrained Ball and Radial Roller Bearings Under Arbitrary Load and Speed Conditions**
A. B. Jones

321 **A Sommerfeld Solution for Finite Bearings With Circumferential Grooves**
J. V. Fedor

327 **Bearing Oil-Ring Performance**
D. C. Lemmon and E. R. Booser

335 **Torque Produced by Misalignment of Hydrodynamic Gas-Lubricated Journal Bearings**
J. S. Ausman

342 **Friction Characteristics of Sliding Surfaces Undergoing Subsurface Plastic Flow**
M. C. Shaw, A. Ber, and P. A. Mamin

347 **Machining Characteristics of Leaded Steel**
F. L. Bagley, Jr., and R. Mennell

360 **A General Flow Graph Technique for the Solution of Multiloop Sampled Systems**
R. Ash, W. H. Kim, and G. M. Kranc

371 **Control System Analysis and Design Via the "Second Method" of Lyapunov, Part I**
R. E. Kalman and J. E. Bertram

394 **Control System Analysis and Design Via the "Second Method" of Lyapunov, Part II**
R. E. Kalman and J. E. Bertram

After a year's experience with the new ASME Transactions quarterlies, it has become necessary to reconsider the assignment of Hydraulic Division papers. Since the Hydraulic Division has been assigned to the *Journal of Basic Engineering*, all hydraulic papers have appeared in this quarterly. However, certain Hydraulic Division papers seem better suited for publication in the *Journal of Engineering for Power*.

Henceforth, papers dealing with prime movers, pumping machinery, compressors, and fluid power systems will, in most cases, be assigned to the *Journal of Engineering for Power*. Papers on water hammer, cavitation, and fluid mechanics will, in general, be assigned to the *Journal of Basic Engineering*.—EDITOR.

EDITORIAL DEPARTMENT

Editor, J. J. JAKLITSCH, JR.
Consulting Editor, LEO BLODGETT
Production, JEAN NORTH
Art Editor, HENRY BALFOUR

OFFICERS OF THE ASME

President, W. L. CISLER
Secretary, O. B. SCHIER, II
Treasurer, E. J. KATES
Assistant Treasurer, H. J. BAUER

PUBLICATIONS COMMITTEE

Chairman, B. G. A. SKROTZKI
HENDLEY N. BLACKMON
R. D. MINDLIN
MARTIN GOLAND
VITO L. SALERNO

Junior Representatives
A. S. GOLDSTEIN
J. W. FOLLANSBEE

REGIONAL ADVISORY BOARD

Region I, H. M. ROGERS, JR.
Region II, GLENN R. FRYLING
Region III, F. J. HEINZE
Region IV, FRANCIS C. SMITH
Region V, H. M. CATHER
Region VI, C. R. EARLE
Region VII, M. S. UMBENHAUER
Region VIII, CARL J. ECKHARDT

Published quarterly by **The American Society of Mechanical Engineers.** Publication office at 20th and Northampton Streets, Easton, Pa. The editorial department is located at the headquarters of the Society, 29 West Thirty-Ninth Street, New York 18, N. Y. Cable address, "Mechaneer," New York. Price $3.00 a copy, $10.00 annually; to members, $1.50 a copy, $5.00 annually. Add 50 cents for postage to all countries outside the United States and Canada. Changes of address must be received at Society headquarters seven weeks before they are to be effective on the mailing list. Please send old as well as new address. . . . By-Law: The Society shall not be responsible for statements or opinions advanced in papers or . . . printed in its publications (B13, Par. 4). . . . Second-class mail privileges authorized at Easton, Penna. . . . This publication is authorized to be mailed at the special rates of postage prescribed by Section 132.122 . . . © **Copyrighted, 1960,** by The American Society of Mechanical Engineers. Reprints from this publication may be made on condition that full credit be given the **Transactions of the ASME, Series D—Journal of Basic Engineering,** and the author, and that date of publication be stated.

TRANSACTIONS OF THE ASME
is indexed by the Engineering Index, Inc.

CONTENTS

401 **Relations Between the Notch Tensile Strength of Cylindrical and Prismatic Specimens of Titanium Alloys and Heat-Treated Steels**
G. Sachs, J. G. Sessler, R. F. Pray, and T. H. Yeh

411 **Nonmetallic Inclusions and Fracture Behavior of Steels**
S. Yukawa

417 **Fracture Mode Transition for a Crack Traversing a Plate**
G. R. Irwin

426 **Use of the Moiré Effect to Measure Plastic Strains**
A. Vinckier and R. Dechaene

435 **A Generalization of Cumulative Damage**
R. M. Mains

441 **Effect of Relaxation on the Behavior of Materials Under Combined Alternating and Static Stress**
F. H. Vitovec

447 **Cyclic Operation of Pressure Piping with γ Heating**
K. R. Merckx

453 **Creep Rupture Tests for Design of High-Pressure Steam Equipment**
E. A. Davis

462 **Steady Creep of a Tube Under Combined Bending and Internal Pressure**
Iain Finnie

465 **Effect of Wall Thickness on Stress-Rupture Life of Tubular Specimens**
J. T. Tucker, Jr., E. E. Coulter, and L. F. Kooistra

477 **Design of Pressurized Cylinders for High-Temperature Applications**
J. F. Traexler

482 **The Response of a Pressure-Sensing System**
R. P. Benedict

489 **Dynamics of Heat Exchangers and Their Models**
Herman Thal-Larsen

TRANSACTIONS OF THE ASME Series A

Journal of
ENGINEERING FOR POWER
Volume 82 • Series A • Number 2

CONTENTS FOR APRIL, 1960

83 Critical Flow Velocities for Collapse of Reactor Parallel-Plate Fuel Assemblies, by D. R. Miller.

96 The Propeller Turbines of the St. Lawrence Power Project, by G. G. Fintak. (59—Hyd-11)

103 Influence of Trailing-Edge Geometry on Hydraulic-Turbine-Blade Vibration Resulting From Vortex Excitation, by Gunnar Heskestad and D. R. Olberts. (59—Hyd-7)

111 Water-Hammer Damage to Oigawa Power Station, by C. C. Bonin. (59—Hyd-8)

120 An Experimental Investigation of Radial Thrust in Centrifugal Pumps, by A. Agostinelli, D. Nobles, and C. R. Mockridge. (59—Hyd-2)

127 Design and Test of Mixed-Flow and Centrifugal Impellers, by J. J. Kramer, W. M. Osborn, and J. T. Hamrick. (59—Hyd-20)

136 Volute Pressure Distribution, Radial Force on the Impeller, and Volute Mixing Losses of a Radial Flow Centrifugal Pump, by H. W. Iversen, R. E. Rolling, and J. J. Carlson. (59—Hyd-10)

145 Relationship of Coal-Ash Viscosity to Chemical Composition, by W. L. Sage and J. B. McIlroy. (59—Fu-4)

156 Pressurization of Granular Solid Fuels, by M. N. Aref. (59—A-150)

TRANSACTIONS OF THE ASME
Series B

Journal of
ENGINEERING FOR INDUSTRY
Volume 82 • Series B • Number 2

CONTENTS FOR MAY, 1960

87 Simplified Approach to Understanding the Suction Problems for Reciprocating Mud Pumps, by W. E. Liljestrand. (59—Pet-23)
93 Prevention of Sulfide-Stress Cracking in Sour-Condensate Well Equipment, by W. J. McGuire and L. W. Vollmer. (59—Pet-36)
103 Overstrain Tests on Thick-Walled Cylinders, by S. M. Jorgensen. (59—Pet-1)
122 An Analytical Study of Drill-String Vibration, by J. J. Bailey and I. Finnie. (59—Pet-15)
129 An Experimental Study of Drill-String Vibration, by I. Finnie and J. J. Bailey. (59—Pet-16)
136 The Theory of the Ideal Design of a Compound Vessel, by S. J. Becker and L. Mollick. (59—A-125)
143 The Strength of Thick Cylinders Subjected to Repeated Internal Pressure, by J. L. M. Morrison, B. Crossland, and J. S. C. Parry. (59—A-167)
154 The Interaction of Textile Structures and High-Speed Textile Processes, by H. M. Morgan. (59—A-115)
159 Effect of Transient Loads on the Refrigerated-Space Temperature of a Cooling Complex, by C. F. Kayan. (59—A-181)
167 Uniform Two-Way Orientation of Plastic Films, by M. O. Longstreth and Turner Alfrey, Jr. (59—A-113)

TRANSACTIONS OF THE ASME
Series C

Journal of
HEAT TRANSFER
Vol. 82 • Series C • No. 2

CONTENTS FOR MAY, 1960

77 Heat Transfer From Transverse and Yawed Cylinders in Continuum, Slip, and Free Molecule Air Flows, by L. V. Baldwin, V. A. Sandborn, and J. C. Laurence. (58—A-202 and 59—HT-5)
87 Magnetohydrodynamic Effects Upon Heat Transfer for Laminar Flow Across a Flat Plate, by R. D. Cess. (59—HT-14)
94 Conical Turbulent Boundary Layer Experiments and a Correlation With Flat Plate Data, by W. S. Bradfield. (59—HT-6)
101 The Influence of Free Stream Turbulence on the Local Heat Transfer From Cylinders, by R. A. Seban (59—HT-3)
108 Heat Transfer to Fully Developed Laminar Flow in a Circular Tube With Arbitrary Circumferential Heat Flux, by W. C. Reynolds. (59—HT-13)
113 Steam Slip—Theoretical Prediction From Momentum Model, by S. Levy. (59—HT-15)
125 Pressure Drop and Heat Transfer in a Duct With Triangular Cross Section, by E. R. G. Eckert and T. F. Irvine, Jr. (59—HT-10)
139 Heat Transfer in Flow Through Rotating Ducts, by C. Y. Kuo, H. T. Iida, J. H. Taylor, and F. Kreith. (59—HT-12)

TECHNICAL BRIEF

152 Comparison of Turbulent Heat-Transfer Results for Uniform Wall Heat Flux and Uniform Wall Temperature, by R. Siegel and E. M. Sparrow.

The Fluids Engineering Laboratory at the University of Michigan

G. V. EDMONSON
Professor, College of Engineering, University of Michigan, Ann Arbor, Mich. Mem. ASME

Although the purpose of this paper is specifically that of describing the Fluids Engineering Laboratory at the University of Michigan, it is clear that an understanding of the faculty philosophy underlying the planning of this unit is a necessary part of the document. This is an age of technological expansion, the rate-of-change of which exceeds any progress the world has hitherto known. The seeking of new knowledge and the application of that knowledge is the work of an ever-increasing number of competent scientists and engineers not only skilled in a technology but equally successful in the art of human understanding and relationships. Institutions of higher education are an integral part of this technological age and, because of this fact, find themselves confronted with the task of foreseeing the educational and research needs of the future. This is an immense task, one which is occupying many of the best minds of this generation. The technological and social progress of the coming generation depends upon the continuing flow of students emanating from our institutions of learning. They must be adequately prepared for the responsibilities which they alone can assume.

Faculty Planning

IT IS THE PURPOSE of any good engineering faculty to plan programs in education which will best prepare the student for his chosen career. The best way as conceived by one group may be different than the best way as conceived by another group, the reason being that the process, in final analysis, involves people, each with special talents and an individual philosophy of life. The programs are, therefore, a blending of those talents and philosophies to a common purpose.

Certainly, the first requirement of any faculty is that of providing an ethical and social environment in which the student may grow morally and socially into an individual acceptable to his fellow student and to those with whom he will one day participate in his technical career. Beyond this, any educational program should be directed toward extending the student's inherent creative and analytical ability. A sound, logical, thought process and a background of fundamental information are essential to the outstanding performance of the professional man in any field of endeavor.

Examination of past and present curriculums and an intelligent evaluation of changing needs of a profession would appear to be reasonable criteria upon which to proceed in the planning of both educational programs and physical facilities. Studies attendant to curriculum changes never cease in an engineering college. An alert faculty is never satisfied with its past performance. Planning for a physical facility, however, cannot change continuously since a decision to proceed with construction is mandatory. It behooves a faculty group to plan well and adequately for a future need that is difficult to decide upon in this technological age.

The Fluids Engineering Laboratory is the result of faculty planning which resulted in the publication of "Faculty Requirements." The "Faculty Requirements" is a written document specifically designed for the guidance of the architects and building designers. It included all the detailed information developed by the faculty concerning functional arrangement and relation, capacity, distribution, special facilities and areas, special requirements, and occupancy. These plans of the faculty have been wholeheartedly supported by the University Administration, the Board of Regents of the University, and the Governor and Legislature of the State of Michigan. When completed, the State can boast of one of the most complete educational laboratories in the nation devoted to the study of the behavior of fluids, one of the very common forms of substances with which the engineer must deal.

In order to understand the reasons for some of the faculty details, it is considered essential that a background be provided. Adequate laboratory facilities have been a real concern in the College of Engineering at the University of Michigan. Adequate total space required to train the numbers of students properly and also relief from obsolescence are matters of utmost concern. In any facilities program, it becomes necessary to have established the use to which any one facility will be put; it is important to know what the faculty visualize in laboratory activity not only for the immediate future but the rather far-distant future; and it is required that the space be developed in order that the cost of building and the operating budgets be kept to a minimum.

An increased activity in engineering college laboratories is to be expected in all advancing areas of technological progress; however, laboratory activity usually lags behind in the educational program. Because of this apparent fact, it becomes necessary to study the historical as well as the future needs for space.

Engineering laboratories are areas in which the research workers and students experience the application of sciences to engineering processes or to specialized technological developments. In the case of the undergraduate student, his laboratory experience is largely confined to his learning the physical phenomena predicted by the basic principles and fundamentals of the profession. The graduate student and the faculty laboratory activity is directed toward the extension of the fundamental knowledge in the several technological areas of development through advanced application of the sciences and mathematics. The space that has heretofore been occupied by laboratories reflects the two functions and is obsolete when it no longer serves both. Complete obsolescence is a fact in many educational institutions.

Early examination of space needs as projected by the several professional departments at the University of Michigan indicated that each had certain laboratory interests which could be satisfied in a common physical plant where height, width, floor loading, and utilities supply were universal needs for their separate labora-

Contributed by the Hydraulic Division and presented at the Hydraulic Conference, Ann Arbor, Mich., April 13–15, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, December 31, 1958. Paper No. 59—Hyd-16.

tory activity. Further, particularly for the undergraduate teaching function, it was immediately apparent that common physical equipment with minor variation frequently satisfied the need of several departments. Although it was obviously impossible because of the weight of numbers to do away with duplication of equipment, certain economies of space and instrumentation, and higher use factor of a kind of equipment appeared possible. The possibility of providing high-capacity utilities, which might become the requirement of any group, was enhanced by giving consideration to distribution of common utilities to all groups within one facility. Finally, it is expected that the entire level of teaching and research in the very important subject, "Behavior of Fluids," will be upgraded through the combined effort, experience, and talents of well-qualified men from the several departments.

The Faculty Study Committee adopted a liberal definition of a fluid, thereby extending its interests not only from the liquid to gaseous phases of substances, but also into the area of solid suspensions in gases and liquids. In those fluid-study areas where special facility requirements were necessary, for instance, combustion studies involving more than normal ventilation requirements and reaction studies where blowout protection for personnel is necessary, it was agreed that a building separate from a general education and research unit should be considered.

It soon became apparent that all faculty groups looked toward high-capacity utilities in different combinations to satisfy their projected research thinking and planning. For instance, Civil Engineering required high water quantities under closely controlled conditions; Chemical Engineering required high water quantities and high quantities of steam, and a modest amount of electric power; Mechanical Engineering required modest amounts of steam, large electric power distribution, and high quantities of water; Engineering Mechanics required a modest amount of water, carefully controlled, and electric power; Aeronautical Engineering required electric power and ample surrounding space of the wind-tunnel area. To satisfy the quantity and the varying utility demands, it was necessary that some of the utilities be generally distributed, and in other cases that a grouping be considered so that the departmental relationship in the laboratory could be maintained.

The utility distribution costs soon indicated that certain types of space needs, namely, classrooms, offices, and small laboratories, likewise sought by the faculty, should be planned in a building other than the Fluids Engineering Laboratory, since the usage of that space did not indicate the need for high-capacity utilities. If, however, this space were provided in the same building, it would be traversed by an unused high-capacity distribution of

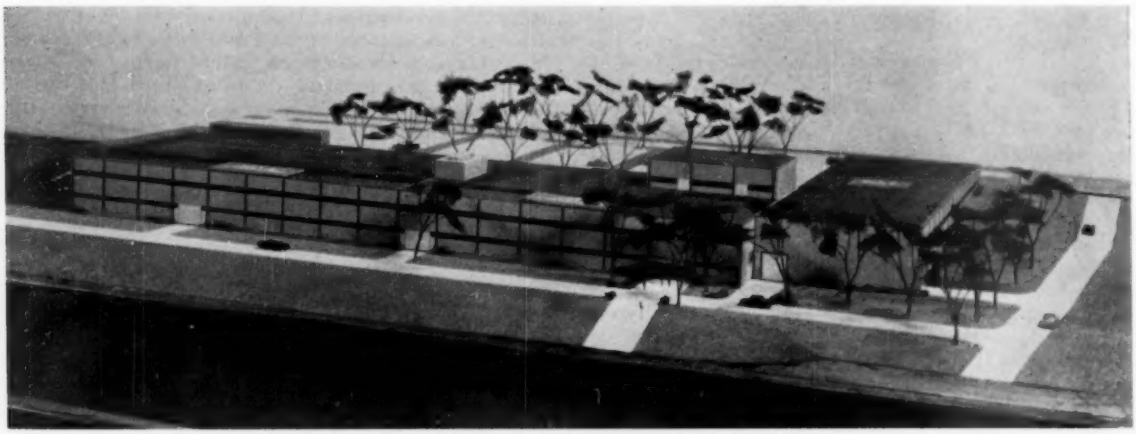


Fig. 1 Fluids Engineering Laboratory model viewed from the front

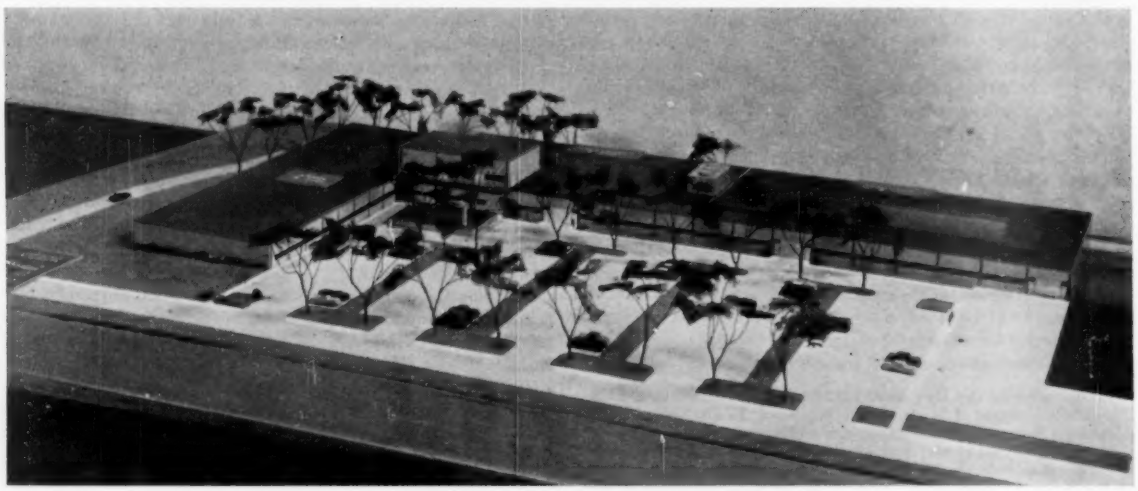


Fig. 2 Fluids Engineering Laboratory model viewed from the rear

those utilities. The cost of distribution of the utilities in the Fluids Engineering Laboratory exceeds the cost of distribution of utilities required by a classroom and office-type of building by approximately four times. The difference in the cost of utilities for the Fluids Engineering Laboratory and a comparable size of classroom building will provide approximately 2½ per cent more space of the classroom type in a total project of approximately 800,000 sq ft. This building program is considered minimum by the Engineering College at the University of Michigan for the future numbers of students who will seek admission.

The Fluids Engineering Laboratory

When the Fluids Engineering Laboratory project is completed, it will provide 120,000 sq ft of laboratory, teaching, and research space. The first portion of the project which provides all the basic utility arrangement is completed. Approximately 52,000 sq ft of laboratory is available in this portion of the project. The complete building is divided into two types of construction. The main portion of the building which is 408 ft long and 128 ft wide, is the usual column-type arrangement: Seventeen bays long and five bays wide and two stories high except in one small portion of the building which is three stories high. The annex, which will be completed with the second portion of the construction, is 230 ft long by 84 ft wide, built of rigid frame construction and one story high. The building basically provides both a large area where normal building column interferences are encountered and a smaller area, the Annex, which is free from vertical columns within the space.

The large portion of the laboratory is oriented east and west with reference to its long dimension and located so that access to either the first or second floor from ground level can be had by foot or vehicle. This flexibility was accomplished by taking advantage of some natural topography. Figs. 1 and 2 are photographs of the laboratory model as viewed from either side of the main building and indicate the accessibility to the first and second floor. The rigid frame portion of the building, the Annex, is adjacent to the east end of the larger portion of the building and is oriented north and south. Access to this portion of the building is gained at the level of the first floor of the main building from either the south or east side.

Building column spacing and ceiling heights were of special concern to the faculty planning committee. Except for the transverse spacing of the building columns in the center of the structure, all columns are spaced on 24-ft centers. The center transverse spacing is 32 ft to provide a clearance for the fluid sump. The net clearance underneath the floor joist on the first and partial third floor is 15 ft; the clearance under the roof joist on the second floor is 12 ft. The net maximum clearance under the rigid frame of the Annex portion of the laboratory is 27 ft. Fig. 3 is a longitudinal section through the main portion of the building which will show the foregoing relationship.

The load capacity of the first floor of the main building and that of the Annex is 400 psf. The load capacity of the second and partial third floor of the main building is 220 psf, respectively. The capacity of the first floor was dictated by the probable load that equipment might impose, the second and third-floor capacity was dictated by the concentrated wheel loading of the fork-lift truck which is the materials-and-equipment-handling device for the complete laboratory. All floors are hard-surface concrete; the office areas are asphalt tiled.

Figs. 4 and 5 show the general plan of the first and second floors, respectively. In order to be able to make installations that exceed the height of the ceiling of the first floor, a 4-ft × 4-ft manhole, mounted flush with the second-floor surface, the periphery of which is enclosed by an open drain trench to the sewer, has been provided in the center of each bay. The manhole covers afford the same load capacity as the second floor thereby avoiding any hazard of break-through. The hard-surfaced floors are trowelled to provide a ½ in. to 10 ft slope to drain from each of the four columns at any bay. In addition to the manholes which could be used for utility connection, four 4-in. flush floor sleeves at the corners of each building column are provided for further utility distribution and communication as may be needed between floors. In the area under the partial third floor, approximately 30 per cent of the floor area is provided with removable plate sections in order that a total height of installation of approximately 45 ft and a maximum of 4 ft in diameter can be accommodated. In a small area of the first floor, the depth of the pump pit No. 6 can be used to gain an additional height of installation to a total of approximately 56 ft.

A high-load-capacity concrete pad 100 ft wide and 72 ft long is provided immediately outside of the high bay area, and on the north side of the main building, for installation of outside-column installations. In order that connections of utilities and instrumentation from within the building to these outside installations can be made, a series of 1-ft × 1-ft portholes are provided in the building wall. In addition to convenience, these portholes avoid the necessity of breaking the vapor barrier which is an integral part of the building walls.

The building walls are made up of 8 in. of brick, 2-in. foam-glass insulation, a membrane vapor barrier, and a 4-in. compressed-block inner wall. The net glass area in the building is less than 20 per cent. All glass is of the double-pane, sealed construction. This construction was specifically directed at the problem of condensation. The interior temperature of the laboratory walls is calculated to be above the dew point on the coldest days of the year.

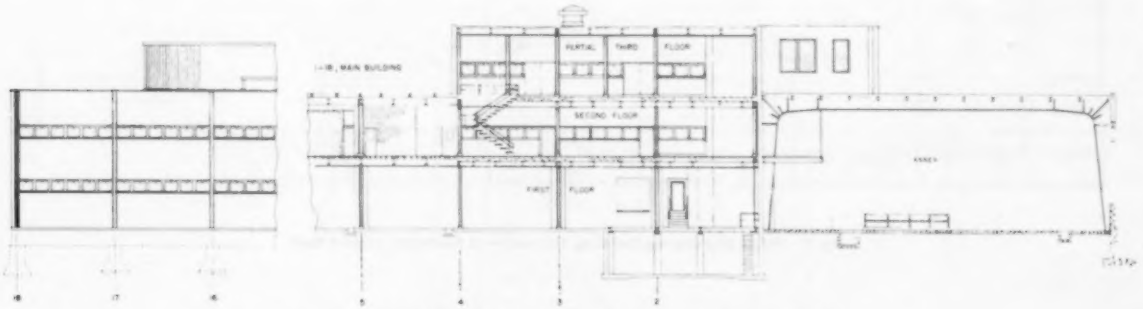


Fig. 3 Longitudinal building section, section viewed from south side

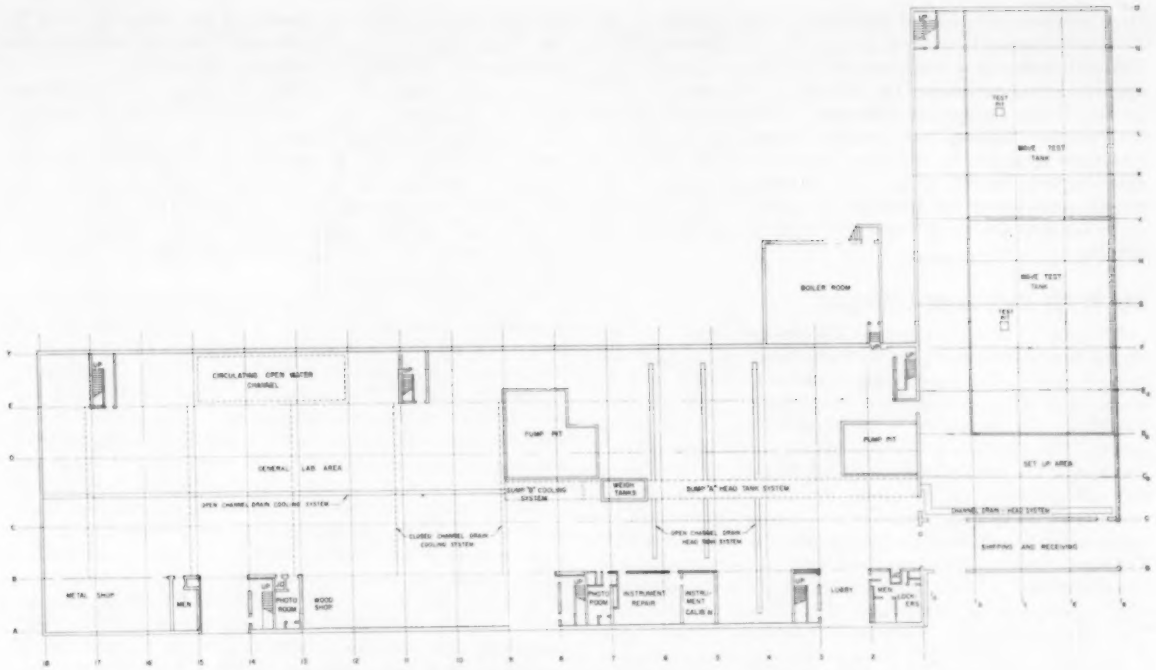


Fig. 4 Fluids Engineering Building, University of Michigan, first floor

Fig. 5 Fluids Engineering Building, University of Michigan, second floor

Building Utilities

Utilities for the building are generally subdivided by function; those that service the building generally and those that are specifically sized and distributed for power equipment used in research and teaching functions.

Service utilities such as telephone outlets, city water for drinking fountains and rest rooms, 117-volt electric distribution, and like services are generally distributed. Four-gang, 30-amp electric receptacles connected to a four-wire, 208-volt distribution affording 117 volts to the neutral wire are placed at each building column. This circuit is separately supplied in order to eliminate voltage fluctuations that might be caused by the starting of a heavy motor on a power circuit. Each outlet is separately fused, thus providing a total of 120 amp at each column. Welder outlets for 480-volt, 3-phase are provided at every second column.

The heavier utilities as required for teaching and research (water, gas, electric, and steam) are distributed in a series of supply loops all hanging from the ceiling of the first floor of the main building.

All the utility services including gas and air supply are available immediately below and to the sides of the manhole openings from first to second floor. This arrangement is shown in Fig. 6(*a*). Except for the head-tank system, it is possible to bring a maximum utility capacity up to the second floor at any bay in the building. In addition to the manhole opening, four 4-in. floor sleeves are arranged at each building column for the purpose of utility distribution. These are shown in Fig. 6(*b*). All utilities are flexibly arranged in order that ready connections can be made to supply research and teaching equipment on either the first, second, or third floor. The supply loops run longitudinally, the length of the loop system being so located laterally that a minimum length of connection can be accomplished from any location in the building.

There are three water systems in the building, two of which are supplied from a sump 192 ft long, 12 ft wide, and 10 ft deep in the shallow end, and a third supplied from city water supply. Additives to reduce rust and scaling are used in the sump-connected systems. One of the water systems is controlled by a head tank located next to the ceiling of the partial third floor approximately 42 ft above the level of the first floor. The head tank is so designed that it provides approximately 140 ft of weir crest adjusted to an accuracy of $^1/_{16}$ in. out of parallel and perpendicular to gravity lines of the earth. The outlet of the head tank is flexibly connected to a 48-in. standpipe extending down into pump pit No. 6. Two supply systems, 8 and 12 in., respectively, are connected to the standpipe and so valved that either or both systems may operate simultaneously. Since only a regulating quantity of water spills over the head-tank weir, the flow of water to the system is progressively from sump to standpipe to system and return to sump. Flanged outlets from the 48-in. standpipe are provided at both the floor and the ceiling levels at each floor. Venturi meters, which can be calibrated against weigh tanks, are arranged in the 12 and 8-in. systems to measure a maximum flow of 7 and 3 cfs, respectively. The 12-in. system is arranged for a flanged outlet at every second bay; the 8-in. system is arranged with two flanged outlets at every bay.

The second water-supply system from the sump is pressurized to operate in conjunction with a cooling tower. This system will deliver 3 cfs at the farthest high point in the building at 60 psi. The pressurized return can be directed through the cooling tower or bypassed directly to the supply sump, depending entirely on the amount of heat absorbed by the water circulating in the system. The system supply is intended for the many heat-transfer functions required in the laboratory. Outlets at every bay, all of which are provided with vertically upward mounted tee connections and flanged caps, are provided from this system. The upward mounted tees are developed to prevent fouling of valves in the supplies to equipment. The cooling towers are designed for a maximum capacity of 30,000,000 Btu per hr.

Both systems are sectionalized in a way that a portion of the system may be cut off for additional connections without shutting down other research activity. Return to sump from either system may be accomplished through transverse drain trenches in the first floor or through a pressurized return.

The weigh tanks, each having a capacity of approximately

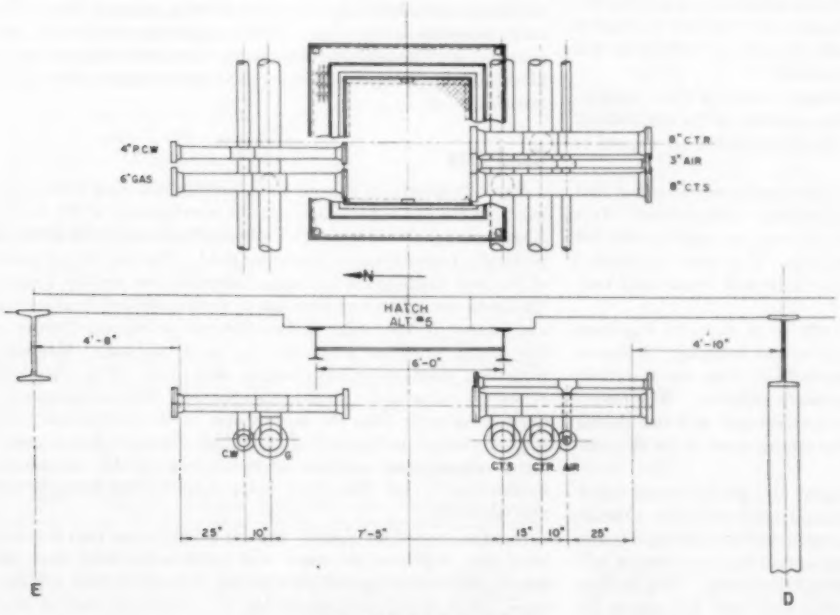


Fig. 6(a) Hatch assembly

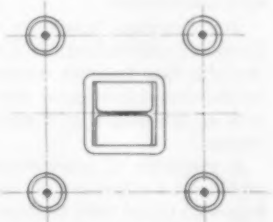


Fig. 6(b) Floor sleeve detail

48,000 lb, are located at the end and immediately over the water sump at the end of sump farthest from the 48-in. standpipe. The weigh tanks are provided with an air-operated diverter in order that calibration quantities exceeding the capacity of one weigh tank may be considered. The tanks may be used to calibrate water quantities from any of the supply systems.

Three-phase electric power at 480 volts is distributed parallel to the fluid supply loops in 600-amp, connected-bus ducts. The bus ducts, as the fluid systems, are hanging from the ceiling of the first floor. A connection at 1-ft intervals through fused plug-ins can be made to supply either the first, second, or third floor. The electric connection between the fused plug-in and the equipment requiring power is largely through flexible rubber-covered drop cable rather than by conduit. Conduit is used in preference to rubber-covered cable only where fixed installations are considered. A 4800-volt circuit is installed in the laboratory for possible future high demand. This circuit is not immediately available through use of plug-in boxes. Special wiring provision from junction boxes will be required to use this circuit.

General distribution of d-c electricity has not been included. Initially, four d-c dynamometers ranging in capacity from 5 to 200 hp have been installed; each, however, supplied by its own motor-generator located in the immediate vicinity of the dynamometer. Where other d-c power is required a rectifier will be installed locally; for instance, 86 kw of 24-volt d-c power is installed in the immediate vicinity of the area where heat-transfer activities are proceeding.

Laboratory Installations

The Fluids Engineering Laboratory has been planned to provide a maximum degree of flexibility for the present and future use of the building. Only those partitions enclosing activities which were likely to be continued, regardless of a possible future requirement of our technology, were included. Such enclosures included a minimum number of offices, computing rooms, a multiphase laboratory, instrument storage rooms, rest rooms, stair wells, photographic rooms, and janitor closets. All other required enclosures will be accomplished using minimum-cost curtain walls, which can be removed and rebuilt as an activity expands or is reduced. In all such changes, the teaching or research activity has available to it the same capacity of utilities as was provided in the space it formerly occupied.

The laboratory includes a minimum number of fixed installations, all of which are of interest to a number of the professional groups. The water weigh tanks described earlier are typical of such installations.

The multiphase laboratory has been partitioned in order that a pressure slightly lower than ambient may be maintained. It is desirable that the solid suspension in gas streams should not be distributed throughout the building. This area is provided with an unusually large number of exhaust and intake roof ventilators.

A gasometer of approximately 600 cu ft capacity has been provided for calibration purposes as well as teaching. A closed-loop wind tunnel with an approximate 24 × 32-in. throat section, likewise required for teaching purposes, is included. The return and entering ducts of the tunnel are vertical and the closing horizontal leg is on the roof. This arrangement is for the purpose of conserving floor space.

A large, low-capacity meteorological and smoke tunnel where air velocities in the order of 5 mph can be maintained in a throat section of 6 × 8 ft is planned to be suspended from the rigid frame of the Annex. Access to the throat section is gained from a balcony which is reached from the second-floor level. This facility will extend the important studies that have been in progress for several years on the distribution of air contaminates due to combustion and chemical processes. It further will enable the extension of studies carried on by the meteorological groups in pollen distribution and other research in which they are actively concerned.

The entire roof section of the main laboratory building is arranged with concrete pedestals which include anchor bolts, the top of which is approximately 8 in. higher than the roof deck and spaced in rectangular array on 8 × 12-ft centers. A roof grating capable of a 75 psf loading is installed on top of the roof pedestals in order that outside laboratory activity may be pursued. This space is particularly attractive for solar-radiation experimental apparatus, many activities in meteorology, and provides access and space around the cooling towers for teaching and research.

The laboratory Annex is designed to provide unobstructed space where wave tanks may be installed. Truck access to this space is provided in order that the large volume of materials required for wave-tank studies may be moved readily. Two 80 × 100-ft 4-ft-deep tanks may be provided by installing sections of tank wall. Either or both tanks may be removed when the space is needed for other activities. These tanks and the accompanying height of ceiling are the necessary facilities to continue the studies of beach erosion, forces on underwater structures by wave action, ship-docking studies, and breakwater research studies that have been so necessary to the special lakes and lake shore problems surrounding the State of Michigan. This facility is certain to take on increased significance with the coming impact of the St. Lawrence Seaway.

The major number of installations in the Fluid Engineering Laboratory are readily movable. No hold-down bolts or machine bases are fixed in the floor. Vibration isolation pads are used throughout. In those cases where the normal center of gravity is higher than considered safe, a concrete base for the equipment is designed and it is set on vibration isolation pads.

Although a plan to provide a space for each graduate student has never been achieved in the Engineering College, such is the plan for the Fluids Engineering Laboratory. In order to achieve a maximum of flexibility and still house instrumentation and provide desk space for the student, a portable type of enclosure which can be set up in the open laboratory immediately adjacent to the student's research equipment is being studied. Several of such enclosures are installed. At the completion of a project, the enclosure may be relocated for a second occupant where he may, likewise, enjoy a degree of privacy and close communication to his research work.

Conclusion

Space flexibility to provide for a rapidly changing technology has been the key consideration in the development of the Fluids Engineering Laboratory. The laboratory is not confined to a particular profession or a particular fluid. The new Department of Nuclear Engineering has many interests that revolve around fluid behavior. The long-standing profession of Civil Engineering and its store of knowledge on fluid flow can be big contributors to the special problems attendant to the atomic age. Magnetic radiation, magneto-hydrodynamics, and many other advanced topics are concerned with gaseous behavior. The continuous reductions in order from the macroscopic to the microscopic tend toward a better understanding of fluid phenomena. It is reasonable to expect that a better understanding of the continuous medium may result from work being done in these areas of advanced study.

It is the hope of the Faculty Planning Committee that the new laboratory will provide space and utilities in order that research, understanding, and the teaching of fluid behavior will continue. The established knowledge is a necessary part of new technological areas where advanced study must continue.

S. RAYNOR
Professor of Mechanical Engineering,
The Technological Institute,
Northwestern University,
Evanston, Ill. Mem. ASME

A. CHARNES
Professor of Applied Mathematics,
The Technological Institute,
Northwestern University,
Evanston, Ill.

Flow Parameters in Hydrostatic Lubrication for Several Bearing Shapes[1]

In the case of hydrostatic lubrication the designer of thrust bearings has to make decisions regarding the shape of pads, location of oil holes, and configuration of oil grooves. In this paper several pad shapes and associated oil inlets were investigated using conformal mapping techniques to obtain the total load-carrying capacity, flow rate of oil, oil-film thickness, pressure and velocity distribution.

The results of these calculations permit the designer to approximate his chosen configuration by computed models in order to estimate the flow parameters.

HYDROSTATIC LUBRICATION is the maintenance of an oil film between two stationary surfaces by pressure produced by an external pump. It is different from hydrodynamic lubrication where the oil film is produced by the relative motion of two surfaces.

Hydrostatic lubrication gives considerable reduction of friction and wear, and is of practical importance in applications when the motion is too slow to produce hydrodynamic lubrication, as in the starting of heavy rotating machinery such as turbines, motor-generators, and synchrocondensers. Hydrostatic lubrication is used also in reversal of reciprocating motion, as in pistons, crossheads, and the like. Elementary theory permits computation of the total load-carrying capacity, the flow rate of oil, and pressure and velocity distribution for the so-called hydrostatic step-bearing, which consists of two circular disks, the oil being supplied through a central hole. Aside from this case, the elementary considerations no longer suffice. Recourse must be had to study and solution of the Navier-Stokes equations in suitable approximation. As will be made clear, it was possible to reduce the general case to determining a conformal mapping.

The object of the present paper is to compute the flow parameters for several new cases in order to permit the designer to better approximate his own configuration by solved cases.

The designs investigated are:

1 Circular plates with central hole
2 Square plates with central hole
3 Elliptical plates with oil groove
4 Infinite strip with central hole

General Considerations

The problem to be investigated is the steady flow between fixed parallel planes. The following assumptions are made: (*a*) That the velocity parallel to the boundary planes is small, (*b*) that the separation between the planes is so small that the velocity component in this direction can be neglected, (*c*) that the pressure is known at inlets and outlets. With these and other (listed later) assumptions, the Navier-Stokes equations can be reduced by neglecting: (*d*) the time dependent and inertial terms,[2] (*e*) the body forces, and (*f*) compressibility terms to:

$$\nabla^2 u = \frac{1}{\mu}\frac{\partial p}{\partial x} \tag{1}$$

$$\nabla^2 v = \frac{1}{\mu}\frac{\partial p}{\partial y} \tag{2}$$

$$\nabla^2 w = \frac{1}{\mu}\frac{\partial p}{\partial z} \tag{3}$$

where

u, v, and w are velocity components in the x, y, and z-directions
μ is the (coefficient of) viscosity, and p is the pressure
$2b$ is the oil-film thickness

If the z-axis is perpendicular to the boundary planes, then $w = 0$.

It follows from Equation (3) that under this condition p is a function of x and y only.

The general continuity equation

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$$

reduces to

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{4}$$

Equation (4) determines that a "stream function" $\psi(x, y, z)$ exists so that

$$u = -\frac{\partial \psi}{\partial y} \tag{5}$$

$$v = \frac{\partial \psi}{\partial x} \tag{6}$$

From this one can infer [1, 2][3] that the pressure $p(x, y)$ is a harmonic function. Using the fact that the solution is unique and that u and v vanish at the boundary planes, it can further be shown that

$$u = \frac{1}{2}(z^2 - b^2)\frac{1}{\mu}\frac{\partial p}{\partial x} = -\frac{\partial \phi}{\partial x} \tag{7}$$

[1] The authors wish to express their appreciation to Mr. T. C. Peng for assistance in carrying out and verifying the computations for the examples in this paper.

Contributed by the Lubrication Division and presented at a joint session with the Machine Design Division at the Semi-Annual Meeting, St. Louis, Mo., June 14–18, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, November 20, 1958. Paper No. 59—SA-61.

[2] This assumption may not be accurate near the immediate neighborhood of inlet hole, where the inertia effects may predominate over viscous effects.

[3] Numbers in brackets designate References at end of paper.

$$v = \frac{1}{2}(z^2 - b^2)\frac{1}{\mu}\frac{\partial p}{\partial y} = -\frac{\partial \phi}{\partial y} \tag{8}$$

The function $\phi(x, y, z) = -\frac{1}{2\mu}(z^2 - b^2)p(x, y)$ evidently serves as a velocity potential for u and v in each *layer* z = constant. Also, in each layer, lines of constant pressure are lines of constant "velocity potential" and vice versa. Determination of the pressure therefore determines the whole flow for this class of problems. The problem of computing the flow parameters is thus reduced to finding a potential (or harmonic) function $p(x, y)$ satisfying the boundary conditions on pressure, a problem mathematically identical to problems encountered in electrostatic fields, steady heat conduction, diffusion studies, and so forth.

Problem 1—Circular Plates

Given a bearing, Fig. 1, consisting of two circular plates with central holes. The pressure in the hole is p_i and is zero at the outer boundaries.

Determine: The pressure distribution p
The total load-carrying capacity P
The flow rate Q required to maintain the film thickness h

The problem is solved by the conformal mapping of fluid flow parallel to the ϕ-axis in the w-plane into a flow in the z-plane (where $z = x + iy$) by means of the mapping function

$$w = -m \ln z = -m \ln \rho\, e^{i\theta} \tag{9}$$

It follows that

$$w = \phi + i\psi = -m(\ln \rho + i\theta)$$

or

$$\phi = -m \ln \rho$$

Lines of constant ϕ correspond to lines of constant pressure,

$$-m \ln \rho = \frac{p}{c}$$

or

$$p = p_i\left(\frac{\ln \rho}{\ln \rho_0}\right) \tag{10}$$

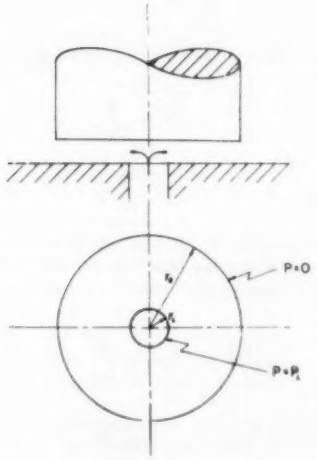


Fig. 1

where

$$\rho = \frac{r}{r_e} \tag{11}$$

$$\rho_0 = \frac{r_i}{r_e} \tag{12}$$

The load-carrying capacity is

$$P = \int_0^{2\pi}\int_{r_i}^{r_e} pr\, dr\, d\theta = -\frac{\pi r_e^2}{\ln \rho_0^2} p_i[1 + \rho_0^2(\ln \rho_0^2 - 1)] \tag{13}$$

and for $\rho_0^2 \ll 1$,

$$P = -\frac{\pi r_e^2}{2 \ln\left(\frac{r_i}{r_e}\right)} p_i = \frac{1}{2 \ln\left(\frac{r_e}{r_i}\right)} p_i A \tag{14}$$

where A is the area of the circular disk. The flow rate is

$$Q = \int_0^{2\pi}\int_{-b}^{+b} v_r\, r\, dz\, d\theta$$

for any r; here v_r is the velocity component in the r-direction

$$v_r = -\frac{\partial \phi}{\partial r} = \frac{1}{2}(z^2 - b^2)\frac{1}{\mu}\frac{\partial p}{\partial r} \tag{15}$$

Upon integration

$$Q = \frac{\pi}{6}\frac{1}{\mu}\frac{h^3}{\ln\left(\frac{r_e}{r_i}\right)} p_i \tag{16}$$

where $h = 2b$ = oil-film thickness.

These results (10), (14), (16) can be obtained directly by elementary methods and are well known. They were computed to show the simplest application of the method developed.

Problem 2—Square Plates

Given a bearing, Fig. 2, consisting of two square disks with

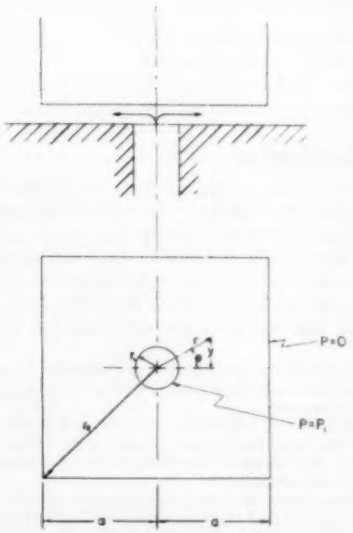


Fig. 2

central holes. The pressure in the hole is p_i and is zero at the outer boundaries.

Determine: The pressure distribution p
The total load-carrying capacity P
The flow rate Q required to maintain the film thickness h

The problem is solved by superposition of a source flow (Problem 1) for an infinite plate, and a secondary flow chosen such that the resulting pressure along the outer boundary of the square vanishes. We shall, for simplicity, however, merely determine the latter in a close approximation.

The mapping function for a source flow is

$$w_1 = -m \ln z = 2c \ln z \tag{17}$$

The mapping function (approximate) for the secondary flow is chosen to have fourfold symmetry

$$w_2 = a_0' + a_4' z^4 + a_8' z^8 \tag{18}$$

where a_0', a_4', a_8' are to be determined so that the pressure vanishes at five chosen points at the outer boundary.

The resulting mapping function is

$$w = 2c' \ln z + a_0' + a_4' z^4 + a_8' z^8 = \phi + i\psi \tag{19}$$

The real part of w, i.e., ϕ, is the velocity potential, and the lines of ϕ = constant correspond to lines of constant pressure in the z-plane.

$$\mathrm{Re}(w) = \ln \rho^2 + a_0 + a_4 \rho^4 \cos 4\theta + a_8 \rho^8 \cos 8\theta = \frac{p}{c'} \tag{20}$$

where

$$\rho = \frac{r}{r_s} \tag{21}$$

At the outer boundary

$$r^2 = \frac{a^2}{\cos^2 \theta} = \frac{r_s^2}{2 \cos^2 \theta} \tag{22}$$

and $p = p_b = 0$

$$\frac{p_b}{c'} = \ln \left(\frac{1}{2 \cos^2 \theta} \right) + a_0 + a_4 \frac{\cos 4\theta}{4 \cos^4 \theta} + a_8 \frac{\cos 8\theta}{16 \cos^8 \theta} \tag{23}$$

Imposing the conditions

$$\left.\begin{aligned} \theta &= 0 & p_b &= 0 \\ \theta &= 22.5 \text{ deg} & p_b &= 0 \\ \theta &= 45 \text{ deg} & p_b &= 0 \end{aligned}\right\} \tag{24}$$

permits the determination of a_0, a_4, and a_8 by solving three simultaneous equations obtained by introducing condition (24) into the equation (23), whence

$$a_0 = 0.5415674 \qquad a_4 = 0.5933672 \qquad a_8 = 0.0518$$

Equation (20) takes the form

$$\frac{p}{c'} = \ln \rho^2 + 0.5415674 + 0.5933672\, \rho^4 \cos 4\theta + 0.0518\, \rho^8 \cos 8\theta \tag{25}$$

Equation (25) yields among other things the lines of constant pressure in polar co-ordinates ρ and θ. By the choice of w_2 the pressure was made zero at the outer boundary of the square at five points along each boundary line, viz., θ = 45 deg, 22.5 deg, 0, −22.5 deg, and −45 deg. By choosing more terms in the expression for w_2 the pressure can be made to vanish at more points on the boundary. The deviation of the pressure from the zero-value along the outer boundary was calculated along the side of the square for the following points, Fig. 3.

$\frac{y_b}{a}$ =	0	0.1	0.2	0.3	0.4	0.5
$\frac{p_b}{c'}$ =	−0.0000004	−0.0001794	−0.0005928	−0.001025	−0.0007445	−0.0000004
$\frac{y_b}{a}$ =	0	0.6	0.7	0.8	0.9	1.0
$\frac{p_b}{c'}$ =	−0.0000004	+0.0011438	+0.0022155	+0.0025856	+0.0018168	+0.000002

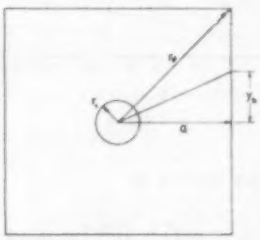


Fig. 3

The maximum deviation from zero $\frac{p_b}{c'} = 0.0025856$ can be expressed in terms of p_i for a given $\frac{r_i}{a} \cong \frac{1}{10}$

$$p_b = \frac{1}{2000} p_i$$

This value can be calculated from Equation (25).

Introducing a new parameter $\beta = \frac{r}{a}$ Equation (25) takes the form

$$\frac{p}{c'} = \ln \frac{\beta^2}{2} + 0.542 + 0.148\beta^4 \cos 4\theta + 0.00324\beta^8 \cos 8\theta \tag{26}$$

Lines of constant pressure were computed by choosing θ constant and computing p as a function of $\beta = \frac{r}{a}$, Fig. 4.

The stream lines (constant value of the stream function) were computed by setting the value of the imaginary part of the complex potential w equal to a constant; $\mathrm{Im}(w)$ = constant = c''

$$\frac{\psi}{c''} = 2\left(\theta - \frac{\pi}{4}\right) + a_4 \beta^4 \sin 4\theta + a_8 \beta^8 \sin 8\theta \tag{27}$$

Solve for

$$\beta^4 = \left(\frac{r}{a}\right)^4 = \frac{-a_4 \sin 4\theta \pm \left(a_4^2 \sin^2 4\theta - 4a_8 \sin 8\theta \left[-\frac{\psi}{c'} + 2\left(\theta - \frac{\pi}{4}\right)\right]\right)^{1/2}}{2a_8 \sin 8\theta} \quad (28)$$

The values of r as a function of θ for constant values of $\frac{\psi}{c'}$ are plotted in Fig. 4.

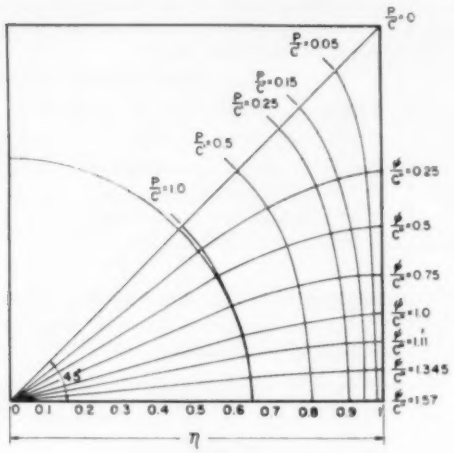


Fig. 4

The load-carrying capacity is

$$P = 8\int_0^{\frac{\pi}{4}} \int_{r_i}^{\frac{a}{\cos\theta}} pr\, dr\, d\theta \quad (29)$$

where

$$\frac{p}{c'} = \ln \frac{\beta^2}{2} + a_0 + a_4\beta^4 \cos 4\theta + a_8\beta^8 \cos 8\theta$$

c' can be evaluated by setting $r = r_i \ll a$ for the inner boundary of the plate where the pressure p is equal to the inlet pressure p_i.

$$c' = \frac{p_i}{\ln \beta_0^2 - 0.15158} \quad (30)$$

Integrating the Equation (29) results in

$$P = \frac{8a^2}{\ln \beta_0^2 - 0.15158} p_i \left[-0.4634 + \frac{\pi}{4}\left(0.458\frac{\beta_0^2}{2} - \frac{\beta_0^2}{2}\ln\frac{\beta_0^2}{2}\right)\right] \quad (31)$$

where

$$\beta_0^2 = \frac{r_i^2}{a^2} \quad (32)$$

For $\beta_0 \ll 1$

$$P = -\frac{3.7012\, a^2}{\ln\left(\frac{r_i^2}{a^2}\right) - 0.15158} p_i = \frac{0.925}{2\ln\left(\frac{r_e}{r_i}\right) - 0.54157} A \quad (33)$$

where A is the area of the square plate.

The flow rate is

$$Q = \int_0^{2\pi} \int_{-b}^{+b} v_r r\, dz\, d\theta$$

where

$$v_r = -\frac{\partial \phi}{\partial r} = \frac{1}{2}(z^2 - b^2)\frac{1}{\mu}\frac{\partial p}{\partial r} \quad (15)$$

For $r_i \ll a$

$$p = c' \ln\left(\frac{\beta_0^2}{2}\right) + a_0 c'$$

and

$$\frac{\partial p}{\partial r} = \frac{2c'}{r} = 2\frac{\left(\ln\frac{\beta^2}{2} + a_0\right)^{-1}}{r} \quad (34)$$

Substituting (34) into (15) yields

$$Q = \frac{\pi}{6}\frac{1}{\mu}\frac{h^3}{\ln\left(\frac{r_e}{r_i}\right) - 0.271} p_i \quad (35)$$

Problem 3

Given a bearing, Fig. 5, consisting of two elliptic plates with an oil groove connecting the two focuses of the ellipse. The pressure in the groove is p_i and the pressure at the outer boundary is zero.

Determine: The pressure distribution p
The total load-carrying capacity P
The flow rate Q required to maintain the film thickness h

The problem is solved by conformal mapping of the source flow (Problem 1) in the w-plane by means of the mapping function

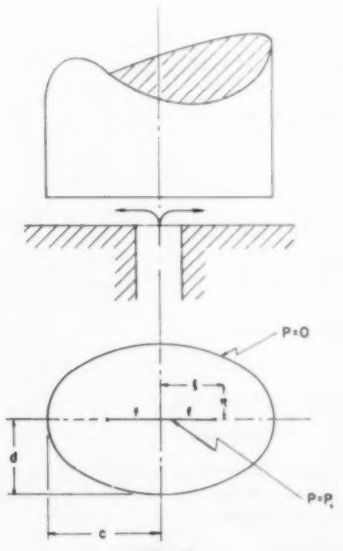


Fig. 5

$$z = \frac{1}{2}\left(\frac{w}{\rho_0} + \frac{\rho_0}{w}\right) \tag{36}$$

This function transforms the inside of the ρ_e circle of the w-plane into the inside of the ellipse in the z-plane. The circle ρ_0 is mapped into the focal slit.

Substituting $w = \rho^{i\theta}$ into the Equation (36) and setting $z = x + iy$ one obtains

$$x = \frac{1}{2}\left(\frac{\rho}{\rho_0} + \frac{\rho_0}{\rho}\right)\cos\theta$$

$$y = \frac{1}{2}\left(\frac{\rho}{\rho_0} - \frac{\rho_0}{\rho}\right)\sin\theta$$

or

$$\frac{x^2}{\frac{1}{4}\left(\frac{\rho}{\rho_0} + \frac{\rho_0}{\rho}\right)^2} + \frac{y^2}{\frac{1}{4}\left(\frac{\rho}{\rho_0} - \frac{\rho_0}{\rho}\right)^2} = 1$$

Changing the co-ordinates to

$$x = \frac{\xi}{f} \qquad y = \frac{\eta}{f}$$

results in

$$\frac{\xi^2}{\frac{f^2}{4}\left(\frac{\rho}{\rho_0} + \frac{\rho_0}{\rho}\right)^2} + \frac{\eta^2}{\frac{f^2}{4}\left(\frac{\rho}{\rho_0} - \frac{\rho_0}{\rho}\right)^2} = 1 \tag{37}$$

For $\rho = \rho_0$, the ellipse degenerates into a slit $-f \leq \xi \leq f, \eta = 0$. If $\rho \gg \rho_0$, the ellipse approaches a circle.

For example, if

$$\frac{\rho}{\rho_0} = 10,$$

then

$$\text{semimajor axis} = 5.05\,f$$

$$\text{semiminor axis} = 4.95\,f$$

The lines of constant ρ correspond to lines of constant pressure in both the w and the z-planes, Fig. 6. The lines of constant pressure are ellipses given by the Equation (37). Introducing the source flow,

$$p = p_i \frac{\ln \rho}{\ln \rho_0} \tag{38}$$

where

$$\left.\begin{aligned} \rho_0 &= \frac{R_0}{R_e} \\ \rho &= \frac{R}{R_e} \\ \rho_e &= \frac{R_e}{R_e} = 1 \end{aligned}\right\} \tag{39}$$

at $\rho = \rho_0$, $p = p_i$ in the slit; and at $\rho = \rho_e$, $p = 0$ at the outer boundary.

The major axis of the ellipse is

$$c = \frac{f}{2}\left(\frac{R_e}{R_0} + \frac{R_0}{R_e}\right) \tag{40}$$

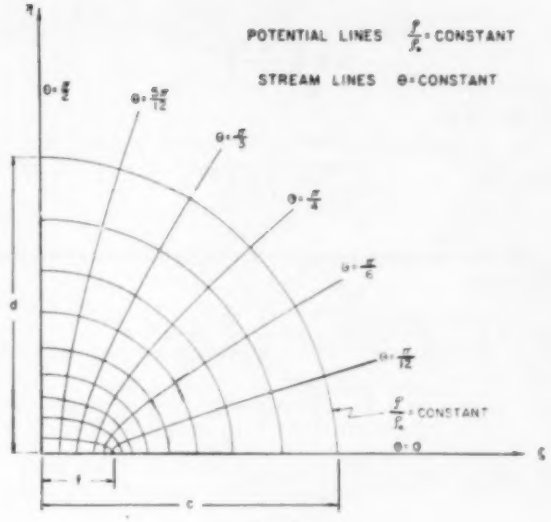


Fig. 6

The minor axis of the ellipse is

$$d = \frac{f}{2}\left(\frac{R_e}{R_0} - \frac{R_0}{R_e}\right) \tag{41}$$

It follows from (40) and (41) that

$$c^2 - d^2 = f^2$$

or that the ellipses are confocal.

The value of ρ_0 can be computed

$$\rho_0 = \frac{R_0}{R_e} = \frac{c - d}{f} = \frac{f}{c + d} \tag{42}$$

For c and d much larger than f

$$\rho_0 \cong \frac{f}{D} \tag{43}$$

where D is equal to the diameter of the circle.

The load-carrying capacity is

$$P = \int\int p\, d\xi\, d\eta = f^2 \int\int p\, dx\, dy \tag{44}$$

integrated over the ellipse.

Using the Jacobian transformation of the area $dx\,dy$ in the z-plane into the area $\rho\,d\rho\,d\theta$ in the w-plane results in

$$P = \int_{\rho_0}^{1}\int_{0}^{2\pi} p f^2 \left|\frac{dz}{dw}\right|^2 \rho\, d\rho\, d\theta \tag{45}$$

$\left|\frac{dz}{dw}\right|^2$ can be computed from (36)

$$\left|\frac{dz}{dw}\right|^2 = \frac{1}{4\rho_0^2}\left(1 + \frac{\rho_0^4}{\rho^4} - 2\frac{\rho_0^2}{\rho^2}\cos 2\theta\right) \tag{46}$$

Substituting (46) and (38) into (45) and integrating, one obtains

$$P = -\frac{\pi f^2}{4 \ln \rho_0^2} p_i \left(\frac{1}{\rho_0} - \rho_0\right)^2 \tag{47}$$

The area of the ellipse A can be introduced, and the relations (42) and (43) substituted into (47) result in

$$P = \frac{1}{2 \ln \left(\frac{c+d}{f} \right)} \frac{d}{c} p_i A \tag{48}$$

or for $c, d \gg f$

$$P = \frac{1}{2 \ln \left(\frac{D}{f} \right)} p_i A \tag{49}$$

The flow rate can be computed considering the outflow from the oil groove

$$Q_z = 2 \int_{-b}^{+b} \int_{-f}^{+f} \frac{1}{2} (z^2 - b^2) \frac{1}{\mu} \frac{\partial p}{\partial \eta} \, d\xi \, dz$$

$$= 2 \int_{-b}^{+b} \int_{-1}^{+1} \frac{1}{2} (z^2 - b^2) \frac{1}{\mu} \frac{\partial p}{\partial y} \, dx \, dz \tag{50}$$

In order to compute $\frac{\partial p}{\partial y}$ notice that

$$p = f(\rho); \quad \frac{\partial p}{\partial y} = \frac{\partial p}{\partial \rho} \frac{\partial \rho}{\partial y}$$

From (38) follows

$$\frac{\partial p}{\partial y} = \frac{p_i}{\ln \rho_0} \frac{\partial \rho}{\partial y} \frac{1}{\rho} \tag{51}$$

$\frac{\partial \rho}{\partial y}$ can be computed by considering

$$\begin{aligned} x &= F(\rho, \theta) \\ y &= G(\rho, \theta) \end{aligned} \quad \text{or} \quad \begin{aligned} F(\rho, \theta) - x &= 0 \\ G(\rho, \theta) - y &= 0 \end{aligned} \tag{52}$$

Differentiate (52) with respect to y

$$F_\rho \frac{\partial \rho}{\partial y} + F_\theta \frac{\partial \theta}{\partial y} = 0$$

$$G_\rho \frac{\partial \rho}{\partial y} + G_\theta \frac{\partial \theta}{\partial y} - 1 = 0$$

Solve for $\frac{\partial \rho}{\partial y}$

$$\frac{\partial \rho}{\partial y} = \frac{-F_\theta}{F_\rho G_\theta - F_\theta G_\rho} \tag{53}$$

Substituting in (53) for

$$F = \frac{1}{2} \left(\frac{\rho}{\rho_0} + \frac{\rho_0}{\rho} \right) \cos \theta$$

$$G = \frac{1}{2} \left(\frac{\rho}{\rho_0} - \frac{\rho_0}{\rho} \right) \sin \theta$$

gives

$$\frac{\partial \rho}{\partial y} = \frac{\rho_0}{\sin \theta} \tag{54}$$

The differential $\frac{\partial p}{\partial y}$ can be computed

$$\left(\frac{\partial p}{\partial y} \right)_{\rho = \rho_0} = p_i \frac{1}{(1 - x^2)^{1/2}} \tag{55}$$

Substituting (55) into (50) results in

$$Q = \frac{\pi}{6} \frac{1}{\mu} \frac{h^3}{\ln \left(\frac{c+d}{f} \right)} p_i \tag{56}$$

or for $c, d \gg f$

$$Q = \frac{\pi}{6} \frac{1}{\mu} \frac{h^3}{\ln \left(\frac{D}{f} \right)} p_i \tag{57}$$

Problem 4

Given a bearing, Fig. 7, consisting of a narrow infinite strip with a central hole, where the pressure is p_i, reducing to zero at the outer boundary.

Determine: The pressure distribution p
The total load-carrying capacity P
The flow rate Q required to maintain the film thickness h

The problem is solved by conformal mapping of the source flow (Problem 1) in the w-plane by means of the mapping function

$$\frac{w - 1}{w + 1} = ie^{iz} \tag{58}$$

This transforms the inside of the unit circle 1, 2, 3, 4 in the w-plane into the inside of the infinite strip 1, 2, 3, 4 in the z-plane.

Substituting $w = \rho e^{i\theta}$ into Equation (58) and setting $z = x + iy$, there follows

$$y = \frac{1}{2} \ln \frac{(\rho^2 + 2\rho \cos \theta + 1)^2}{\rho^4 - 2\rho^2 \cos 2\theta + 1}$$

$$\tan x = \frac{1 - \rho^2}{2\rho \sin \theta}$$

A small inlet hole of radius $\rho_0 \ll 1$ corresponds to a circular hole in the z-plane

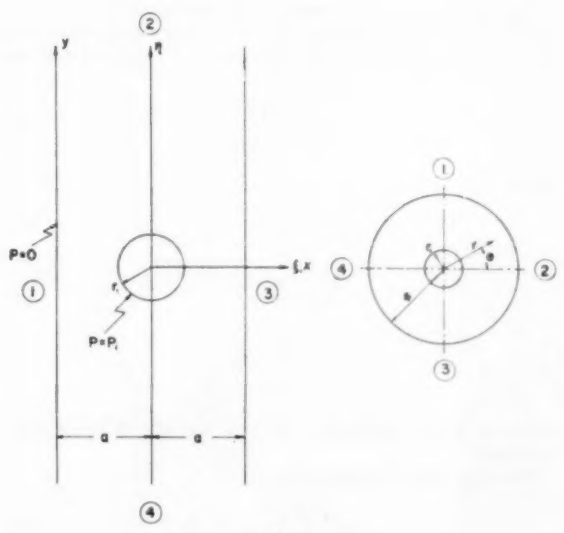


Fig. 7

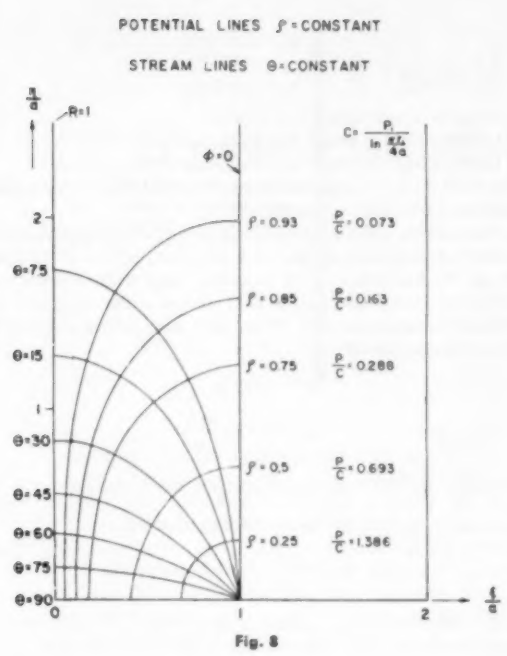


Fig. 8

$$\left(\frac{y}{2\rho_0}\right)^2 + \left(\frac{x - \frac{\pi}{2}}{2\rho_0}\right)^2 = 1$$

Changing the co-ordinates to

$$x = \frac{\pi}{2} + \frac{\xi}{a}\frac{\pi}{2}$$

$$y = \frac{\eta}{a}\frac{\pi}{2}$$

there follows

$$\left.\begin{aligned} \frac{\xi}{a} &= \frac{2}{\pi}\tan^{-1}\left(\frac{1-\rho^2}{2\rho\sin\theta}\right) - 1 \\ \frac{\eta}{a} &= \frac{1}{\pi}\ln\frac{(\rho^2 + 2\rho\cos\theta + 1)^2}{\rho^4 - 2\rho\cos\theta + 1} \end{aligned}\right\} \tag{59}$$

The lines of constant ρ correspond to lines of constant pressure in the w-plane; they are also constant-pressure lines in the z-plane, Fig. 8. The lines of constant θ correspond to streamlines, Fig. 8.

In the w-plane

$$p = p_i \frac{\ln\left(\frac{R}{R_e}\right)}{\ln\left(\frac{R_i}{R_e}\right)} \tag{60}$$

or

$$p = p_i \frac{\ln\rho}{\ln\rho_0} \tag{61}$$

For $\rho_0 \ll 1$ in the z-plane, there corresponds a circle with the radius r_i (inlet radius).

$$\left(\frac{\xi_i}{a}\right)^2 + \left(\frac{\eta_i}{a}\right)^2 = \left(\frac{4\rho_0}{\pi}\right)^2 = \left(\frac{r_i}{a}\right)^2 = r_0^2$$

$$r_i = \frac{4a}{\pi}\rho_0 \tag{62}$$

By elimination of ρ_0 from (62) and (61) there follows

$$p = c\ln\rho \tag{63}$$

where

$$c = \frac{p_i}{\ln\left(\frac{\pi}{4}\frac{r_i}{a}\right)} \tag{64}$$

Total load-carrying capacity

$$P = \int_{-a}^{+a}\int_{-\infty}^{+\infty} p\, d\xi\, d\eta = \frac{4a^2}{\pi^2}\int_0^{\pi}\int_{-\infty}^{+\infty} p\, dx\, dy$$

Using the Jacobian transformation of the area $dx\,dy$ in the z-plane into the area $\rho\,d\rho\,d\theta$ in the w-plane yields

$$P = \frac{4a^2}{\pi^2}\int_{\rho_0}^{1}\int_0^{2\pi} p\left|\frac{dz}{dw}\right|^2 \rho\, d\rho\, d\theta \tag{65}$$

$\left|\frac{dz}{dw}\right|^2$ can be computed from (58)

$$\left|\frac{dz}{dw}\right|^2 = \frac{4}{\rho^4 - 2\rho^2\cos 2\theta + 1} \tag{66}$$

Substituting (66), (63), (62) into (65) and integrating, one obtains

$$P = \frac{16a^2}{\pi\ln\rho_0^2}\, p_i\left[-\frac{\pi^2}{8} + \rho_0^2(1 - \ln\rho_0^2)\right] \tag{67}$$

for

$$\rho_0 = \frac{1}{10},\ \rho_0^2(1 - \ln\rho_0^2) = 0.05605,$$

and for

$$\rho_0 \ll 1, \qquad P = \frac{\pi a^2}{\ln\left(\frac{4a}{\pi r_i}\right)}\, p_i \tag{68}$$

The load-carrying capacity of the infinite strip is approximately twice that of a circular disk whose diameter equals the width of the strip.

The flow rate is

$$Q = \int_{-b}^{+b}\int_0^{2\pi}\frac{1}{2}(z^2 - b^2)\frac{1}{\mu}\frac{\partial p}{\partial r}\, r\, d\psi\, dz \tag{69}$$

For small r

$$p = p_i \frac{\ln\left(\frac{\pi}{4}\frac{r}{a}\right)}{\ln\left(\frac{\pi}{4}\frac{r_i}{a}\right)}$$

and

$$\frac{\partial p}{\partial r} = \frac{p_i}{\ln\left(\frac{\pi}{4}\frac{r_i}{a}\right)}\frac{1}{r} \tag{70}$$

Substituting (70) into (69) and integrating results in

$$Q = \frac{\pi}{6}\,\frac{1}{\mu}\,\frac{h^3}{\ln\left(\frac{4}{\pi}\,\frac{a}{r_i}\right)}\,p_i$$

Summary of Results

Circular bearing, Fig. 1

$$P = \frac{1}{2\ln\left(\frac{r_o}{r_i}\right)}\,p_i A \qquad Q = \frac{\pi}{6\mu}\,\frac{h^3}{\ln\left(\frac{r_o}{r_i}\right)}\,p_i$$

Square bearing, Fig. 2

$$P = \frac{0.925}{2\ln\left(\frac{r_o}{r_i}\right) - 0.54157}\,p_i A \qquad Q = \frac{\pi}{6\mu}\,\frac{h^3}{\ln\left(\frac{r_o}{r_i}\right) - 0.271}\,p_i$$

Elliptic bearing, Fig. 5

$$P = \frac{1}{2\ln\left(\frac{c+d}{f}\right)}\,\frac{d}{c}\,p_i A \qquad Q = \frac{\pi}{6\mu}\,\frac{h^3}{\ln\left(\frac{c+d}{f}\right)}\,p_i$$

Infinite strip bearing, Fig. 8

$$P = \frac{\pi a^2}{\ln\left(\frac{4a}{\pi r_i}\right)}\,p_i \qquad Q = \frac{\pi}{6\mu}\,\frac{h^3}{\ln\left(\frac{4a}{\pi r_i}\right)}\,p_i$$

References

1 L. V. Streeter, "Fluid Dynamics," McGraw-Hill Book Company, Inc., New York, N. Y., 1948, pp. 277–282.

2 H. J. S. Hele-Shaw, "Investigation of the Nature of the Surface Resistance of Water and of Stream-Line Motion Under Certain Experimental Conditions," *Trans. Institution of Naval Architects*, vol. 40, 1898.

DISCUSSION

Richard Elwell[4]

The authors have presented a unique method of analyzing hydrostatic thrust bearings. Usefulness of the technique appears to be unlimited, and its presentation comes at a time when there is a great deal of interest in this type of bearing. Up to now, bearings of such odd shapes as ellipses have been solved almost exclusively on electric analog field plotters.

In the derivation of Equation (13) in this paper, the term $\pi p_i r_i^2$ has been omitted. This represents the portion of the load carried by the pocket of the simple circular bearing. The equation should be:

$$P = \frac{-\pi r_o^2}{\ln \rho_0^2}\,p_i\,[1 - \rho_0^2] \qquad (71)$$

A derivation of this may be found in [3].[5] Equation (71) will reduce to (14) for the special case in which $\rho_0^2 << 1$.

To illustrate the error introduced, consider the following bearing, tested in the course of a project which will be the subject of a forthcoming paper [4] on hydrostatic bearings:

$p_i = 120$ psig
$r_i = 0.375$ in.
$r_o = 0.625$ in.

Measured load capacity = 87.3 lb
Load capacity calculated by Equation (71) = 92.2 lb
Load capacity calculated by Equation (13) = 39.1 lb

The term $\pi p_i r_i^2$ for this case amounts to 53.1 lb, which is a major portion of the bearing load capacity.

The authors' solutions for the square plate bearing with central round hole may be compared to a previous solution of this shape by A. M. Loeb [5]. This geometry, and many others, were solved by an electric analog field plotter. The load and flow from [5] for a square plate 20 in. on a side, with a one inch diameter center hole, is:

$$P = 60.1\,p_i$$

$$Q = 0.129\,\frac{p_i h^3}{\mu}$$

The solutions for these quantities from this paper are:

$$P = 53.7\,p_i \text{ (from Equation 31)}$$

$$Q = 0.171\,\frac{p_i h^3}{\mu} \text{ (from Equation 35)}$$

It is seen that load capacity agrees fairly well, but there is a marked divergence in the two flow predictions, with the present paper predicting a full 32 per cent more flow than the electric analog solution.

References

3 D. D. Fuller, "Theory and Practice of Lubrication for Engineers," John Wiley & Sons, New York, N. Y., 1956, pp. 85–89.

4 R. C. Elwell and B. Sternlicht, "Theoretical and Experimental Analysis of Hydrostatic Thrust Bearings," ASME Paper No. 59—A-152.

5 A. M. Loeb, "The Determination of the Characteristics of Hydrostatic Bearings Through the Use of the Electric Analog Field Plotter," *Trans. ASLE*, vol. 1, no. 1, April, 1958, Chicago, Ill.

Authors' Closure

Mr. Elwell's discussion and indication of the excellent electric analog work of Mr. Loeb is greatly appreciated since it affords an opportunity to clarify an important aspect of our paper.

Our illustrative examples *all* refer to cases wherein the internal hole is of very small radius so that the contribution of the oil within it to the load capacity is negligible. As may be noted, in no example has the "pressure-times-area" on the internal hole contribution to the load capacity been added in. Perhaps the basic source of confusion is the fact that the mapping for the first example (when one adds in the pressure contribution of the hole) holds good regardless of the size of the internal hole.

When the hole is of finite size, the authors' method is still valid, but more complicated conformal mappings may be required. E.g., for the case of the square plate with circular hole the approximate "secondary flow" would be determined by requiring its contribution to the pressure to vanish at certain points on the boundary of the hole as well as to make the *total* pressure zero at certain points on the boundary of the square.

In any case, error in the pressure at any point introduced by the approximate mapping can always be easily estimated because the maximum error must occur on the boundary, i.e., either on the internal hole or the boundary of the pad. Thus also systematic development of this type of approximate mapping can lead to a computer method for determining the flow for arbitrary bearing shapes, the main elements of computation involved being solution of a system of simultaneous linear equations.

[4] Bearing and Lubricant Center, General Engineering Laboratory, General Electric Co., Schenectady, N. Y. Assoc. Mem. ASME.

[5] Numbers in brackets designate References at end of discussion.

A General Method for Correlating Labyrinth-Seal Leak-Rate Data

F. E. HEFFNER
Senior Research Engineer,
Research Laboratories,
General Motors Corporation,
Warren, Mich. Assoc. Mem. ASME

The proper selection of an optimum labyrinth seal for a given application depends upon an accurate estimate of the seal leak rate. Suitable accuracy, particularly for straight labyrinth seals, can only be obtained from actual seal tests. Testing a complete series of seals is time-consuming and expensive. A method for correlating test data is presented which allows calculation of the leak rates for an entire family of labyrinth seals on the basis of tests of only two characteristic seals. Leak rates predicted by the method are within 3 per cent of the basic test data.

SELECTION of the optimum labyrinth seal for any given application is highly dependent upon the ability of the seal designer to predict leak rates through the various possible seal designs. Theoretical leakage predictions are not usually accurate enough for careful design decisions, and it has become necessary to modify the theoretical estimates by empirical factors obtained from actual seal tests. Furthermore, maximum usefulness is obtained from these tests only when general relationships between seal leak rates and physical seal-design parameters can be expressed.

General methods of estimating the leakage of "staggered" labyrinths are available in the literature [1, 2, 3, 4],[1] while data upon which accurate estimates of "straight" labyrinth leakage can be based are less extensive and are more closely limited to exact seal configurations which have been tested [3, 7, 8]. Likewise, correlations between empirical data and physical seal-design parameters are more widely applicable for staggered seals than they are for straight seals.

An experimental study of the flow characteristics of straight labyrinth seals of small radial clearance was made with the intention of extending available design data into the small-clearance range by means of one of the existing correlation methods. Analysis of the test data, however, showed that known general relationships for other absolute dimensional ranges were not applicable with radial clearances smaller than 0.010 in., and it was necessary to evolve a new system for expressing relationships between seal designs and leak rates.

[1] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the joint ASLE-ASME Lubrication Conference, New York, N. Y., October 20-22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, July 6, 1959. Paper No. 59—Lub-7.

The method of correlation which was developed for the small-clearance range, however, may be applied to labyrinth seals of any design. It allows prediction of the leak rates of a whole family of labyrinth seals on the basis of data which may be obtained from two tests on components typical of the seal family of interest. In certain instances one or both of these tests may be eliminated and in their stead theoretical relationships or existing data may be used. The accuracy of the leak rates predicted by the method given will be confirmed within 3 per cent by test. The primary purpose of this paper is to discuss the method of correlation. In doing this it is also necessary to present some test data and specific curves based upon these data. It must be emphasized that the correlation method is applicable to either straight or staggered labyrinth seals with blades of any size, shape, or spacing. However, the test data used to demonstrate correlation method must be applied only to straight labyrinth seals within the explicit dimensional limits for which they are given.

Leak-Rate Analysis

Analysis of the leak rates through labyrinth seals is most easily approached by reasoning analogous to that which describes the flow through single orifices. The basic equation for the leak rate through the seal may then have the form of Equation (1) which can be made identical to the familiar orifice equation if the proper pressure ratio function is introduced.

$$M = \alpha A \phi P \left(\frac{g}{RT}\right)^{1/2} \tag{1}$$

Effective use of this equation to express the leak rate through a labyrinth seal depends upon the proper expression of the coefficient α and the proper definition of the seal-pressure-ratio function ϕ.

In the present analysis, α is defined as the contraction coefficient of the blades in the seal, and it is determined by experi-

Nomenclature

A = gas-flow area, sq in.
D_1, D_2 = clearance space, ID and OD, respectively, in.
G = mass leak-rate velocity, $lb_M/(sec)(in.^2)$
g = gravitational constant, $(lb_M)(in.)/(lb_F)(sec^2)$
h = radial clearance, in.
k = gas specific-heat ratio (dimensionless)
M = mass leak rate, lb_M per sec
N = number of blades
N_{Re} = Reynolds number (dimensionless), $N_{Re} = 2Gh/\mu$
P = absolute pressure at seal inlet, lb_F per sq in.
P_a = absolute pressure at seal outlet, lb_F per sq in.
R = gas constant, $(lb_F)(in.)/(lb_M)(deg\ R)$
r = pressure ratio = P_a/P
r_{cr} = critical pressure ratio
T = absolute temperature at seal inlet, deg R
t = blade thickness, in.
W = specific leak rate, $(lb_M)(deg\ R^{1/2})/(sec)(lb_F)$, $W = G(T^{1/2})/P$
α = contraction coefficient
ϕ = pressure-ratio function
μ = viscosity, $lb_M/(in.)(sec)$

mental measurements of leak rates through a single-blade seal. Numerical values of contraction coefficients are calculated from single-blade test data where

$$M = \alpha AP \left(\frac{g}{RT}\right)^{1/2} \left(\frac{2k}{k-1}\right)^{1/2} \frac{1}{r^{\frac{k}{k}}} \left(1 - r^{\frac{k-1}{k}}\right)^{1/2} \quad (2)$$

Correlation between single-blade test data and the blades in a multiple-blade seal is based upon Reynolds number which may be expressed in terms of a specific leak rate as

$$N_{\mathrm{Re}} = \left[\frac{W}{\alpha}\right] \left[\frac{2Ph}{\mu(T)^{1/2}}\right] \alpha \quad (3)$$

The seals which are the subject of the present study are uniformly constructed with the blade dimensions, radial clearances, and blade spacings invariant throughout a given seal. Thus the mass velocity through the seal is constant, the clearance is the same for all of the blades, and the viscosity of the air as it passes through the seal is nearly constant, so that Reynolds number for a single blade in the seal is the same as the Reynolds number of the entire seal.

The pressure-ratio function used in the present analysis is that derived by Egli [3]

$$\phi = \left(\frac{1 - r^2}{N + \frac{2}{k} \ln \frac{1}{r}}\right)^{1/2} \quad (4)$$

Egli's derivation contains several assumptions and approximations, only two of which ever differ enough from actual conditions in a straight labyrinth seal to cause variation between the theoretical and experimental conditions. First, Egli postulated that the velocity from the expansion at one blade must be dissipated in the interblade space, so that the velocity of approach at the succeeding blade is very small. Second, in order to perform some mathematical operations, Egli assumed that the seal would have a "large number" of blades. There is no direct theoretical way to deduce how many blades this "large number" has to be, but it is to be expected that the derived pressure-ratio function will represent flow conditions accurately only in seals composed of some minimum number of blades.

With the parameters thus defined, it is convenient to use the specific leak to eliminate complexities caused by varying initial conditions and to rewrite the basic equation as

$$\frac{W}{\alpha} = \phi \left(\frac{g}{R}\right)^{1/2} \quad (5)$$

Labyrinth-Seal Critical Flow

The similarity between a labyrinth seal and a single orifice includes a critical flow condition of a maximum gas velocity which is uninfluenced by further reductions in downstream pressure.

In a labyrinth seal the static pressure decreases from blade to blade as the gas expands through the successive stages. Since the mass flow is constant throughout the seal, and since the stagnation temperature also remains essentially constant throughout the seal; it can be shown that the maximum gas velocity will occur at the blade farthest downstream. The seal flow becomes critical when this maximum velocity reaches the sonic velocity for the flowing gas. Thus, an ideal labyrinth will reach its critical over-all pressure ratio when the static pressure ratio at the last blade reaches the known critical value for the particular gas being used.

This critical flow condition of the seal may be theoretically defined by the seal-pressure-ratio function, Equation (4) which becomes maximum at a certain "critical" over-all pressure for the respective number of blades. A comparison between this theoretical criterion and experimentally determined critical pressure ratios is shown in Fig. 1. The solid line is the locus of the calculated critical pressure ratios, and the individual points show the experimentally determined over-all pressure ratio at which the static pressure ratio across the last blade of each seal became equal to the critical value for air. Data for seals of four blade spacings are shown.

There is some variation between theory and experiment, but the data points are generally within a 0.05 pressure ratio of the calculated curve. The significance of this much variation is best assessed by a consideration of Fig. 2 where complete curves of pressure-ratio functions for several numbers of blades are plotted. It can be seen that the value of the function changes very little for over-all pressure-ratio changes of 0.05 near the maxima. Thus the condition of seal criticality can be indicated by the occurrence of the maxima of the respective pressure-ratio functions.

The labyrinth-seal pressure-ratio function is thus applied in the same manner as the single-orifice pressure-ratio function by postulating that it retains its maximum value for all pressure ratios below that at which the maxima occur. The diagonal dashed line in Fig. 2 shows the maximum values of ϕ for all numbers of blades. It is easily shown that ϕ_{max} is equal to $r_{cr}(k)^{1/2}$.

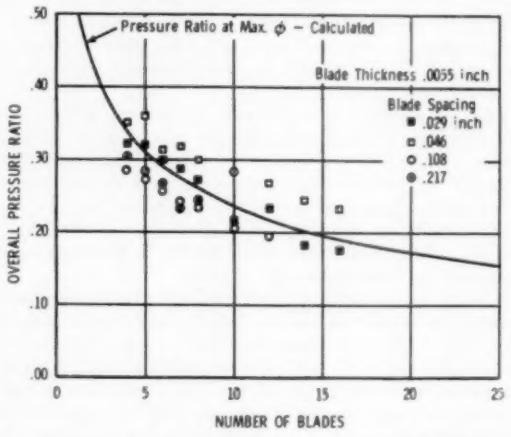


Fig. 1 Theoretical and experimental critical pressure ratios, radial clearance = 0.008 inch

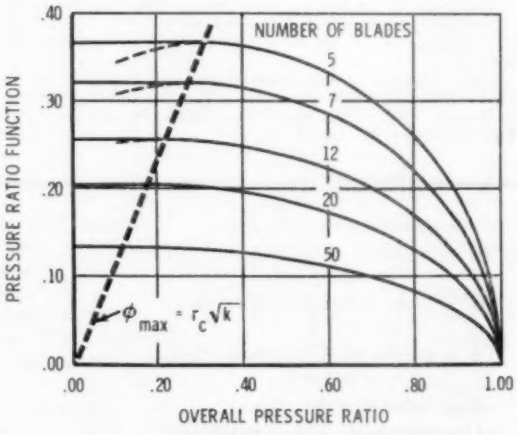


Fig. 2 Labyrinth-seal pressure-ratio functions

Experimental Method

The leak-rate data referred to in the subsequent sections of the paper were obtained by a simple and direct test method. Compressed air was passed in series through a calibrated orifice flowmeter and the test seal, and the leak rate determined from the orifice-meter readings in the customary manner.

The tests were limited to uniform, straight labyrinth seals of radial clearances less than 0.011 in. Only rectangular geomtries were studied. In particular, the upstream and downstream sides of the blades were parallel and normal to the direction of the gas flow; and the steel blade edges were ground to maintain corners as sharp and free from imperfections as possible. The data presented in this paper are for stationary seals only. Fig. 3 is an enlarged scale drawing of the arrangement of the blades and housing for a typical test combination. Significant geometrical nomenclature is shown in the sketch.

Test seals were constructed by assembling suitably sized blades and spacers on a hub to form the bladed seal section. These blades, spacers, and the hub are shown in Fig. 4. This assembly was then inserted into a cylindrical housing to form the labyrinth seal

It is evident that this method of seal construction, while suitably versatile, had the limitation that allowances and tolerances of the component parts would tend to change the seal dimensions slightly. In order to evaluate the effects of these manufacturing variations, seal build-ups and leak tests were repeated periodically during the program to check the reproducibility of the data. In the seal tests thus checked, all of the test points were within three per cent of a mean curve through the data. On the basis of these duplicate tests, the accuracy of the test data which are presented in this paper has been established at ±3 per cent. For the purposes of these labyrinth-seal tests, the calibrated orifice meters which measured the leak rates can be assumed to have had no error.

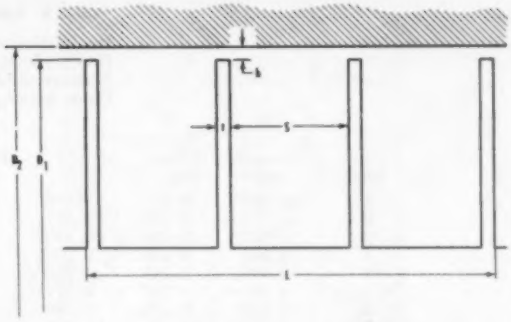


Fig. 3 Geometry and nomenclature sketch

Fig. 4 Hub, blades, and spacers

Construction of Correlation Curves

Two sets of data are required to construct the basic curves from which leak rates through a whole family of seals may be calculated.

First, it is necessary to obtain numerical values of the contraction coefficients for the blades of interest from leak-rate tests on a single blade. These may be plotted against Reynolds number as shown by the solid line in Fig. 5.

Second, it is necessary to determine experimentally the leak rate through a multiple-blade seal of the blade size and spacing of interest. The method of constructing the second correlation curve from these test data is outlined in Table 1.

The data in columns 1, 2, 3, and 4 of the table may be determined directly from multiblade test data. The contraction coefficients in column 5 are taken from the single-blade test results shown as the solid line in Fig. 5. Column 6 may then be calculated. Equation (5) suggests plotting values of W/α as ordinates against the respective values of ϕ in column 3. These data from Table 1 are shown by circular symbols in Fig. 6.

The experimental data show a nearly linear relationship between variables below the critical pressure ratio; but W/α becomes multivalued, and hence indefinite, when the pressure ratio becomes smaller than the critical ratio. The curves of Figs. 5 and 6 therefore define the leak rate of the seal in the manner suggested by Equation (5) only for pressure ratios above the critical.

The following expedient was employed to resolve this ambiguity at less-than-critical pressure ratios. The curve established for the lower values of the pressure-ratio function was extended as shown by the dashed portion in Fig. 6 so that ordinates corresponding to the abscissas of the last four data points are indicated by the two crosses on the curve. It will be seen that the last three points which have a common value of the pressure-ratio function now also have a common value of W/α. The ordinates indicated by these crosses were then recorded in Table 1, column 7, opposite their respective pressure-ratio functions; and new values of the contraction coefficients were computed from columns 7 and 2. The new contraction coefficients are given in column 8, and they have then been plotted in Fig. 5 with their smoothed values indicated by the dashed line.

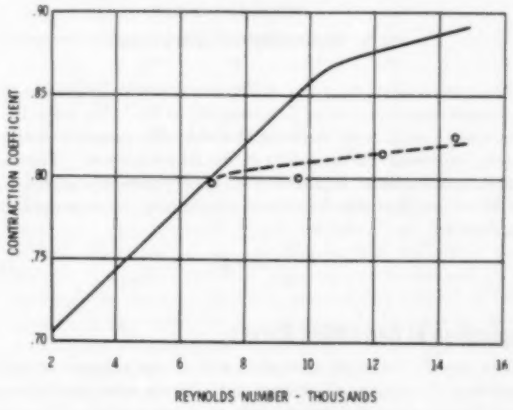


Fig. 5 Contraction-coefficient-curve example

Table 1 Labyrinth-seal test data

Blade thickness 0.0055 in.
Radial clearance 0.008 in.
Number of blades 10
Blade spacing 0.108 in.

Run	1 Pressure ratio	2 Specific leak rate, W	3 ϕ	4 Reynolds number	5 Single blade α	6 $\frac{W}{\alpha}$	7 $\left(\frac{W}{\alpha}\right)'$	8 α'
1	0.833	0.127	0.173	1473	0.691	0.185		
2	0.711	0.166	0.217	2247	0.707	0.235		
3	0.622	0.188	0.240	2906	0.720	0.261		
4	0.551	0.203	0.253	3552	0.732	0.278		
5	0.497	0.215	0.262	4161	0.745	0.288		
6	0.330	0.240	0.277	6991	0.798	0.301	0.304	0.789
7	0.247	0.248	0.280	9669	0.853	0.291	0.308	0.802
8	0.198	0.251	0.280	12226	0.880	0.285	0.308	0.815
9	0.165	0.254	0.280	14841	0.893	0.284	0.308	0.825

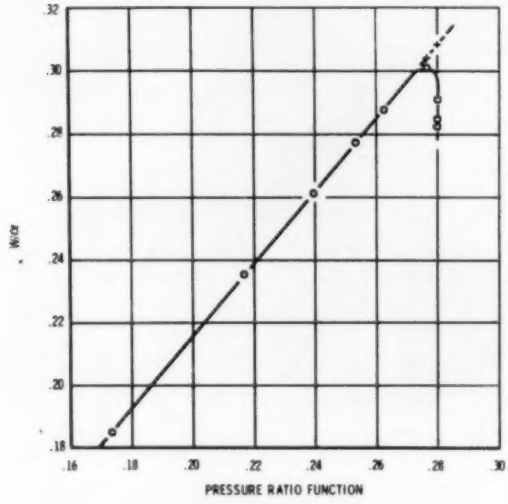


Fig. 6 Seal-correlation-curve example

Also, the slope of the curve in Fig. 6 is numerically greater than the theoretical constant of Equation (5) [0.7767 (lb_M)(deg $R^{1/2}$)/(lb_F)-(sec)], so it must be concluded that this equation does not exactly represent the operation of the labyrinth seal. The more general statement of Equation (5a) is therefore introduced with the intention that this functional relationship be expressed only graphically.

$$\frac{W}{\alpha} = f(\phi) \qquad (5a)$$

Application of Correlation Curves

The specific leak rate through a seal of any number of blades operating at any over-all pressure ratio can be calculated directly from the equations and curves of the previous sections. Equation (5a) as expressed by the curve of Fig. 6 may be used to establish the value of W/α when the pressure-ratio function has been evaluated from Equation (4). This value of W/α can be substituted into Equation (3) with the particular values of the other operating-condition variables to obtain an equation expressing Reynolds number as a function of the contraction coefficient α. Another independent relationship between Reynolds number and the contraction coefficient is expressed by the curve of Fig. 5. The common solution to these two equations establishes the value of the contraction coefficient; and the value of the specific leak rate can be found when W/α and α are known.

Table 2 Recommended minimum numbers of blades

Radial clearance, in.	Blade spacing, in.	Minimum number of blades
0.008	0.029	14
	0.0465	14
	0.108	10
	0.217	7
	0.326	5
0.005	0.029	14
	0.0465	14
	0.108	10
	0.217	8
	0.326	6

This procedure is demonstrated in detail in the Appendix. The example is based upon the data of Figs. 7 and 8 which are from actual tests of seals with 0.008-in. radial clearance and 0.0055-in-thick blades at the various blade spacings shown. Since several spacings are considered in each chart, care must be exercised to ensure the use of the contraction-coefficient curve and leak-rate correlation curve for the spacing of interest.

Discussion of Correlation Method

It has been noted that the pressure-ratio function which is the basis of the correlation curves can be expected to produce discrepancies between theory and experiment for seals of too few blades or when there is velocity carry-over between the blades. The limitations imposed by these two conditions are discussed first.

A demonstration of the blade-number effect is given in Fig. 9. Seal-correlation curves are plotted using data from tests on seals which have too few blades. The upper family of curves for 2, 3, 6, and 10 blades at 0.029-in. spacing shows that the function W/α tends to become too small as the number of blades decreases. The "seal-correlation curves" in this figure are portions of the exact curves of Fig. 7. The data of Fig. 7 are for various numbers of blades above the allowable minima, and that figure shows that the correlation with respect to blade numbers is good for higher numbers of blades.

Further examination of Fig. 9 shows that, as the blade spacing increases, the allowable minimum number of blades tends to decrease. It will be seen that at 0.217-in. spacing the data for only three blades follows the design curve quite closely until the maximum value of the pressure-ratio function is approached. At the larger spacings, 0.651 and 1.302 in., data from tests on two-blade seals approach the theoretical line of Equation (5) very closely.

Empirical analyses like those of Fig. 9 led to the recommended minimum numbers of blades given in Table 2. It will be seen that these minima are functions of both the radial clearance and the blade spacing. These data may be extrapolated without loss

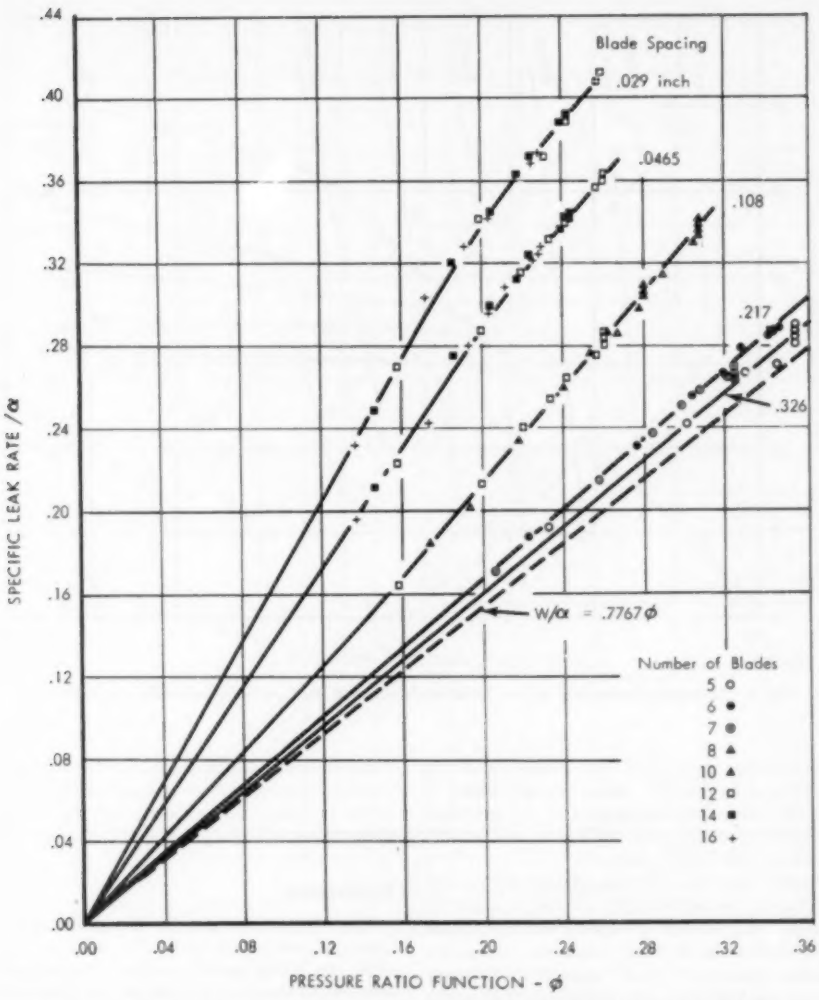


Fig. 7 Seal-correlation curves, blade thickness = 0.0055 in., radial clearance = 0.008 in.

of accuracy to the extent that the minima recommended for a spacing of 0.326 in. may be applied to seals of larger spacings as well.

The velocity-carry-over effect is indicated by increased slopes of the leak-rate curves for the smaller clearances in Fig. 7 where the data are from seals with enough blades to eliminate the blade-number error. It is to be expected that the velocity effect will be greater at small spacing and that it will decrease as more inter-blade space is provided. The curves of Fig. 7 show that when the spacing is 0.326 in., the leak rate is only about 4 per cent greater than theoretical. From this it may be concluded that the line of Equation (5) may be used to calculate the leakage of seals with blade spacings greater than 0.326 in.

The trends shown in the contraction-coefficient curves of Fig. 8 indicate the same conclusion. The apparent contraction coefficients in the multiblade seals approach the actual coefficients for the single blade as the blade spacing increases. This suggests that the gas in the seal nears the zero approach velocity of the single-blade test as the spacing increases. The conclusion that a spacing of 0.326 in. is sufficient to allow use of the single-blade data for the seal-contraction coefficient also comes directly from the method of construction of the correlation curve. For this large spacing the apparent contraction coefficients in column 8 Table 1, are found to equal the actual single-blade coefficients throughout the entire Reynolds-number range.

This discussion of the restrictions imposed upon the system by the basic differences between the theoretical labyrinth seal and the practical labyrinth seal completes the description of the correlation method and establishes its limitations except for a definite statement as to the accuracy of the leak-rate calculations which are based upon the method. If the contraction-coefficient curve is plotted for the exact blade geometry of interest—and this need not necessarily be the rectangular geometry of the present data—and if the correlation curve is constructed from tests on a seal of the same blade geometry with the blade spacing of interest and with more blades than the appropriate minimum given in Table 2; the leak rate through a similar seal of any number of blades above the stated minimum may be calculated to the same accuracy as the experimental data. This accuracy estimate is entirely empirical, and it is based upon data from about 200 test runs at various blade sizes and spacings. When the data were plotted in the manner of Fig. 7, the points were always within 3 per cent of the mean curve. A 3 per cent deviation had been established as the variability of the experi-

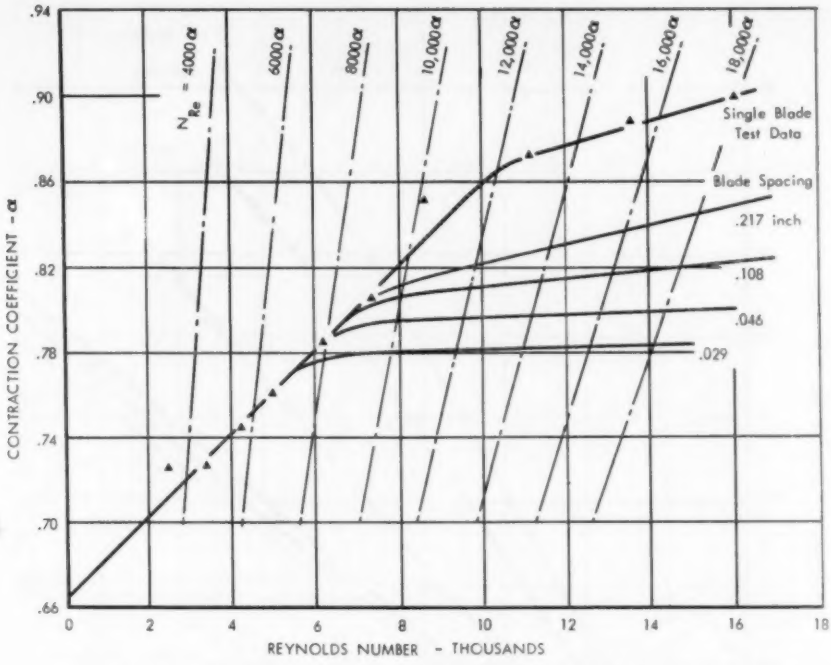


Fig. 8 Contraction-coefficient curves, blade thickness = 0.0055 in., radial clearance = 0.008 in.

ments by means previously mentioned. Extrapolation of the data to larger numbers of blades than those actually tested is allowed because of the theoretical foundation of the pressure-ratio function ϕ, the evidence presented in Fig. 1, and studies like those shown in comparing Figs. 7 and 9.

The correlation method may also be extrapolated with respect to seal diameter. That is, the method may be used to calculate the leak rates through seals of diameters different from those which are tested to obtain the data on which the curves are based. The seal radial clearance is usually so small with respect to the circumference of the seal blade that changing the diameter of the seal while maintaining the same blade dimensions and spacing will change only the area of the leak path. This area change, of course, has no effect on the specific leak rate which is the basis of the correlation method.

In some instances it may be desirable to estimate the leakage through a straight-labyrinth-seal design under circumstances which do not warrant construction of the test apparatus required to obtain the high accuracy of the procedure outlined in this report. In this event, contraction-coefficient data probably can be obtained from the literature since a large number of single-blade tests have been reported [4, 5, 7, 9]. Fig. 10 has been included to supplement these data. Extrapolations of contraction coefficients on the basis of thickness-to-clearance ratios are likely to be as much as 15 to 30 per cent in error, however. This is particularly true in the small clearance range, as the data of Fig. 10 indicate.

The theoretical-correlation curve of Equation (5) may be used for seals with sufficiently large blade spacing or for staggered labyrinth seals. In the absence of additional experimental data from which new correlation curves for particular seals can be drawn, the curves of Fig. 7 are the best estimates available for small blade spacings. If these curves are used for clearances and blade thicknesses other than those for which they are drawn, but for clearances less than about 0.010 in. and blades less than 0.020 in. thick, they will probably estimate within 10 per cent of the correct leak rate provided that the correct contraction-coefficient curve is used. No estimate of the error of extrapolating these curves to larger clearances has been made.

Conclusions

Theoretical estimates of the leak rates through straight labyrinth seals are of very little value in careful seal design because the actual seal operating conditions are appreciably different from those which must be postulated in order to solve the theoretical problem. Accurate estimates of leak rates must be based upon tests of the exact seal design of interest.

A method of correlating test data is presented which allows the calculation of leak rates through an entire family of seals from data which can be obtained from only two tests. If the method is properly applied, the accuracy of the estimated leak rates will be the same as the accuracy of the test data from the required two tests. This may be of the order of 3 per cent as compared to the 20 to 100 per cent errors which might result from purely theoretical calculations.

The method may be applied to either straight or staggered labyrinth seals of any blade shape, size, or spacing; and the required two tests need not be on full-size seals.

The method may be extended to making estimates of greater than purely theoretical accuracy in instances where a test program is not warranted. Under these circumstances, the resulting estimates will probably not be within the 3 per cent accuracy, but the method achieves the maximum of accuracy from whatever pertinent experimental data may be available.

References

1 A. Stodola, "Steam and Gas Turbines," translated by Louis C. Loewenstein, sixth edition, McGraw-Hill Book Company, Inc., New York, N. Y., 1927, vol. 2, pp. 189–194.

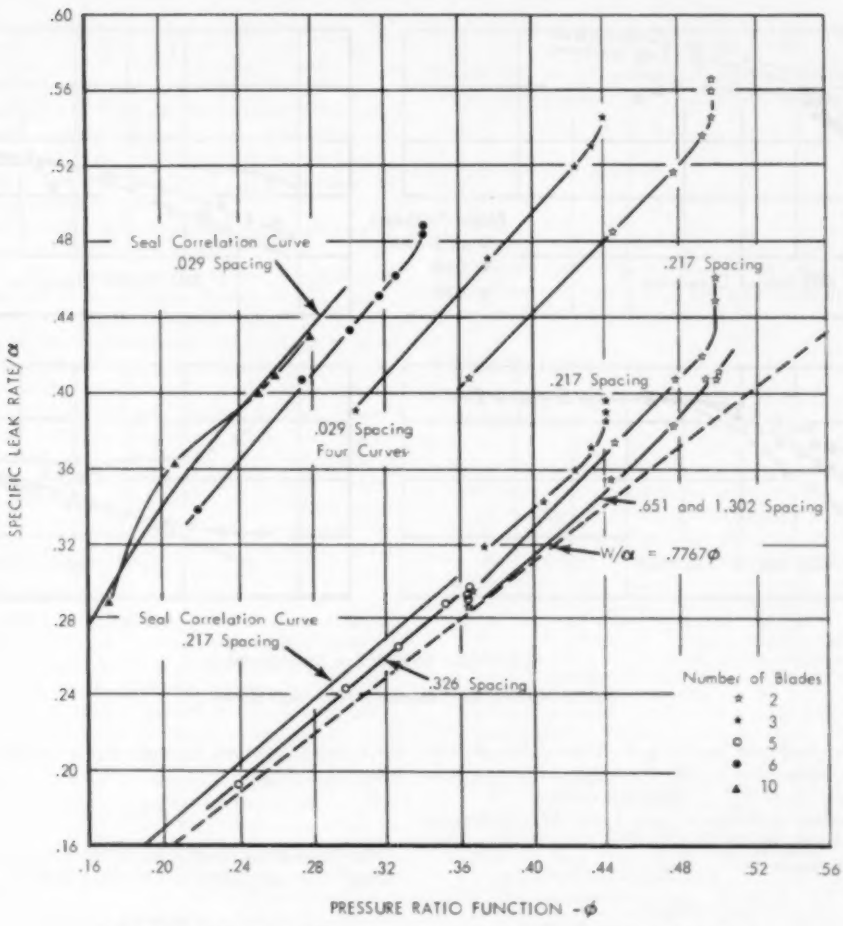


Fig. 9 Effect of number of blades on seal correlation, radial clearance = 0.008 in.

2 H. M. Martin, "Steam Leakage in Dummies of the Ljungstrom Type," *Engineering*, January 3, 1919, pp. 1–3.

3 Adolph Egli, "The Leakage of Steam Through Labyrinth Seals," TRANS. ASME, vol. 57, 1935, pp. 115–122.

4 W. J. Kearton and T. H. Keh, "Leakage of Air Through Labyrinth Glands of Staggered Type," *Proceedings of the Institution of Mechanical Engineers*, vol. 166, 1952, pp. 180–195.

5 W. J. Kearton, "The Flow of Air Through Radial Labyrinth Glands," *Chartered Mechanical Engineer*, vol. 2, 1955, pp. 109–111.

6 W. Zabriskie and B. Sternlicht, "Labyrinth-Seal Leakage Analysis," TRANS. ASME, Series D, JOURNAL OF BASIC ENGINEERING, vol. 81, 1959, pp. 332–340.

7 Jan Jerie, "Flow Through Straight-Through Labyrinth Seals," Proceedings of the Seventh International Congress for Applied Mechanics, 1948, vol. 2, pp. 70–82.

8 J. S. Jones, discussion of Kearton and Keh, op. cit., pp. 189–190.

9 E. W. Snow, discussion of Kearton and Keh, op. cit., pp. 190–191.

10 K. J. Bell and O. P. Bergelin, "Flow Through Annular Orifices," TRANS. ASME, vol. 79, 1957, pp. 593–599.

APPENDIX

Sample calculation of leak rate from correlation curve and contraction coefficient curve.

1 Establish seal-design and operation conditions:

Radial clearance	0.008 in.
Blade spacing	0.217 in.
Number of blades	7
Supply pressure	25.26 psia
Supply temperature	528 R
Over-all pressure ratio	0.5507
Air viscosity	1.05×10^{-6} lb$_M$/(in.)(sec)

2 Determine pressure-ratio function from Equation (4). Carefully establish whether pressure is greater or less than critical value. If ratio is less than critical, use appropriate value of ϕ_{max} (see Fig. 2).

$$\phi = 0.29787 < \text{critical}$$

3 Determine value of W/α from seal correlation curve for proper spacing (Fig. 7).

$$\frac{W}{\alpha} = 0.2520$$

4 Solve for numerical factor of Equation (3) using W/α from step 3 and proper values of variables in general term.

$$N_{Re} = (0.2520)\frac{(2)(25.26)(0.008)}{(1.05 \times 10^{-6})(528)^{1/2}}\alpha$$

$$N_{Re} = 4222\,\alpha$$

5 Sketch the straight line representing the equation of step 4

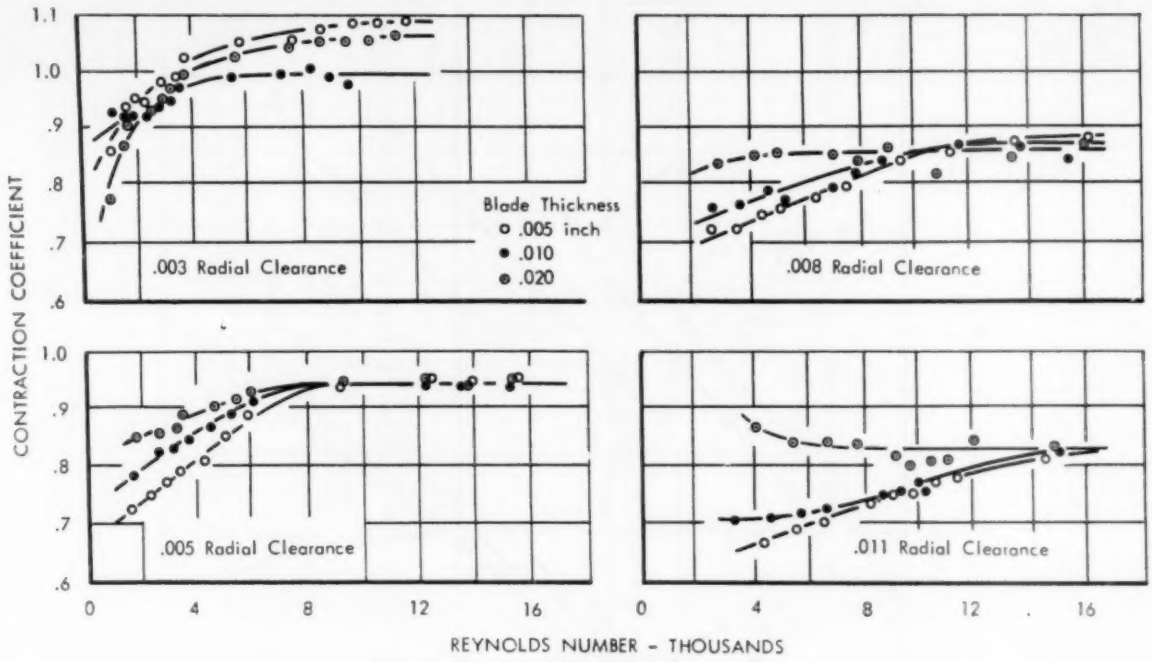


Fig. 10 Contraction-coefficient chart—single blades

on the contraction-coefficient chart, Fig. 8. Typical lines plotted in Fig. 8 may be used as guides. The line required from step 4 is slightly to the right of the $N_{Re} = 4000\alpha$-line shown.

6 The intersection of the line of step 5 and the contraction-coefficient curve for the proper blade spacing is the common solution for α which is required.

$$\alpha = 0.722$$

also

$$N_{Re} = 3010$$

7 The specific leak rate can be found from the value of W/α of step 3 and the value of α of step 6.

$$W = \left(\frac{W}{\alpha}\right)(\alpha)$$

$$W = (0.2520)(0.722)$$

$$W = 0.1819$$

8 The values of the leak rate and the mass velocity can be calculated directly from the specific leak rate and the respective definitions of these quantities.

DISCUSSION

A. Egli[2]

The author's procedure may be summarized as follows:

Introducing a slightly different definition of specific leak rate

$$W^* = W\sqrt{\frac{R}{g}}, \qquad (6)$$

the author started out by representing his tested leakage by the following equation

$$W^* = \alpha\Phi \qquad (7)$$

The contraction coefficient α he determines from a single "blade" test and plots it as a function of Reynolds number

$$\alpha = \alpha(N_{Re})_{\text{single test}} \qquad (8)$$

as shown in his Fig. 5.

The function $\Phi = \Phi(r; N)$ is computed from his equation (4) depending on pressure ratio r across the whole labyrinth and the number of stages N.

In an attempt to correlate equation (7) with test results he runs into difficulties forcing him to two changes.

First, he changes the tested relation (8) to

$$\alpha^* = \alpha^*(N_{Re};\ \text{geometry}) \qquad (8a)$$

Second, he changes equation (7) to

$$W^* = \alpha^* f(\phi) \qquad (9)$$

Both relations are represented graphically:

(8a) in his Fig. 8

(9) in his Figs. 7 and 9

According to the author's statements this method of representation correlates about 200 test runs within 3 per cent. While this result seems to justify a computation method I still feel obliged to criticize this analysis as not rationally founded.

Let us consider the contraction coefficient which represents the "vena contracta" of the fluid jet emerging from each constriction in the labyrinth. In general the jet contraction will depend on three influences:

1 The constriction geometry, represented in the case investigated by the two ratios

[2] Engineering Research and Advanced Product Study Office, Ford Motor Company, Dearborn, Mich. Mem. ASME.

$$\left(\frac{S}{h}\right); \left(\frac{t}{h}\right)$$

2 The compressibility of the fluid, conveniently represented by the pressure ratio r across the constriction of the labyrinth.

3 The viscosity of the fluid, conveniently represented by an appropriately defined Reynolds number N_{Re}, for instance, by equation (3) of the author.

With the author's test rig it appears entirely feasible to properly obtain separately the three influences on α.

Reynolds number, for instance, can be varied (independently from r and seal geometry) by changing the size of the system. In the present case it will be perfectly good enough to vary the dimensions s, h, t while keeping their ratios constant. A range of system sizes of 4 to 1 should be entirely feasible (the cylinder diameter, of course, need not be changed). A kinematic viscosity variation is conveniently obtained by varying the test pressure level (leaving pressure ratios unaffected). A range of at least 2.5 to 1 should be obtained. Thus a variation of N_{Re} in a range of 10 to 1 could be accomplished. Such tests for a single constriction are straightforward. The program and analysis for multistage labyrinths is a little more tricky. One would limit the study at first to possibly three constrictions, just to get an opportunity to obtain the effect of carry-over of velocity energy from one constriction to the next.

Consulting published data of effect of Reynolds number on orifice flow coefficients one would, however, suspect that Reynolds number is not an important factor in labyrinth leakage predictions. In fact, it is the very purpose of a labyrinth to obtain resistance to flow without having to depend on the viscosity of the fluid. If the design situation is such (very narrow clearance, very many constrictions) that really low Reynolds numbers (in the order of 1000 or less) can be attained then it is usually proper to change over to a bushing type or surface friction type seal with much reduced leakage.

Compressibility obviously affects the vena contracta through a sharp edged orifice or slot as has been observed by many experimenters. The effect in fact is such as to even wipe out the critical pressure ratio point in the mass flow versus pressure ratio curve. The discusser strongly suspects that the variation of α shown in the author's Table 1 is primarily due to the compressibility effect and only relatively little of it is caused by fluid viscosity.

Construction geometry affects α of a single constriction as is known from work with orifice flow measuring techniques. In the straight through type arrangement of the multistage labyrinth the amount of velocity energy carried in from the previous constriction will depend largely on the spacing to clearance ratio s/h.

The pressure ratio function, such as ϕ used by the author, is supposed to express the thermodynamic functioning of the labyrinth which involves two processes taking place at each constriction: Production of a high velocity fluid jet and reconversion of a part of its energy back into potential energy.

The author computed ϕ from equation (4) taken from Egli's paper. This equation is certainly not accurate enough once the pressure ratio across one of the constrictions (the last one) drops below about 0.8. And it could not be considered rational to rearrange the discrepancy between test and computation as a viscosity effect.

Concluding this discussion I should like to propose that the tests made by the author could be, quite rationally, expressed in the following manner:

$$W^* = \alpha C \phi \tag{10}$$

where the contraction coefficient α is a function of the seal geometry and the pressure ratio

$$\alpha = \alpha \text{ (geometry, } r) \tag{11}$$

The effect of the seal spacing arrangement is represented by C as a function of geometry and number of constrictions

$$C = C \text{ (geometry, } N).^{5} \tag{12}$$

The pressure ratio function

$$\phi = (r;\ N) \tag{13}$$

to be fairly realistic may have to be built up graphically. Reynolds number has been left out of this representation as it is the discussers' opinion that the viscosity effect is too *academic* a refinement in labyrinth calculations.

William P. Foster[6]

The author's method should materially reduce the amount of seal testing required and expedite evaluation of new configurations.

A number of similar air leakage tests on static straight pass labyrinth seals have been run in our laboratory. These tests confirmed that, in general, close tooth spacing is inefficient.

We find, however, that tooth height is also a factor which affects flow and that there is apparently an optimum tooth spacing to height ratio. For example, a four tooth straight pass seal with a 0.2-in. tooth spacing and a 0.25-in height was better than

[5] Reference might be made to a recent analytical computation of the carry-over effect by K. H. Groddeck, Hannover. A summary of his thesis is published in "*Forschung auf dem gebiete des Ingenieurwesens,*" 1957, no. 5, pp. 183–185.

[6] Engineer, Bearings, Lubrication & Seal Dev. Unit, SAED—Mechanical Systems, General Electric Company, Lynn, Mass.

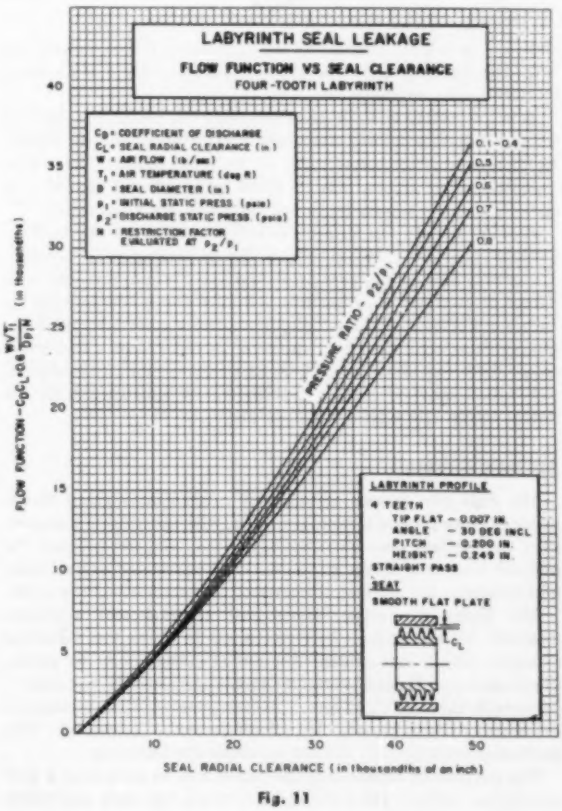


Fig. 11

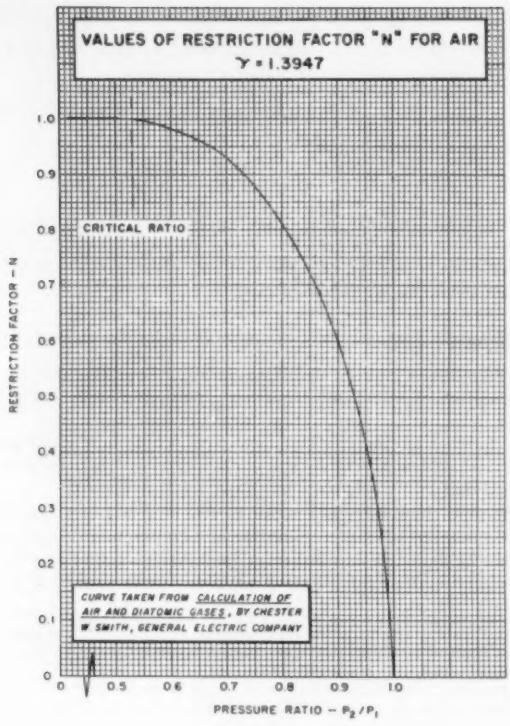


Fig. 12

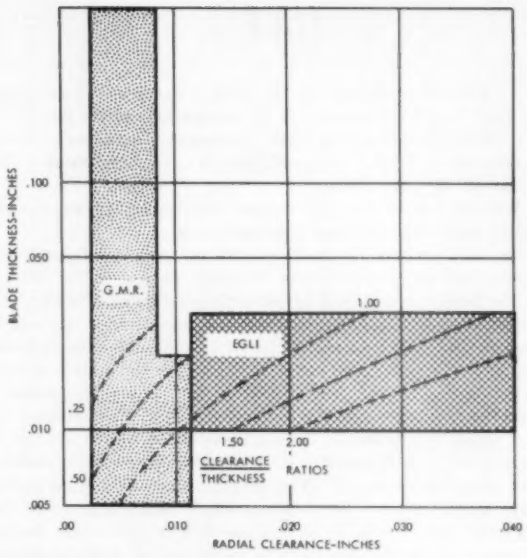


Fig. 13 Experimental clearance and thickness ranges

a 0.15 or a 0.4-in. height. A 0.1-in. spacing 0.2-in. height seal leaked appreciably more than the apparently optimum 0.2 spacing 0.25-in. height combination.

It is interesting to note that increasing the tooth spacing from 0.2 to 0.4 in. and holding the height at 0.25 in. caused the leakage to increase slightly below 0.020 in. radial clearance. This is in violation of the general rule on tooth spacing but it does tend to show that an optimum tooth spacing to height ratio exists.

It also confirms the author's conclusions that, for the best accuracy, each seal configuration should be evaluated by testing.

Our own leakage data are presented as a series of pressure ratio curves plotted versus radial seal clearance and a flow function as shown in Fig. 11. The flow function includes mass flow, upstream temperature, seal diameter, upstream pressure, and a restriction factor read from the curve in Fig. 12. Leakage for any configuration tested may be quickly determined by substituting in the flow function and solving for mass flow.

Author's Closure

Mr. Egli paraphrases the explicit relationships given in the paper in more general terms; concedes, with apparent reluctance, that the method does correlate test data; objects because the method does not conform to his own particular rationalism; and reiterates the method of reference [3] as a preferable system.

Mr. Egli's discussion emphasized rational and academic methods, so the author feels obligated to state the following syllogism which was evident but not emphasized in the paper. *Major premise:* Known methods did not correlate the test data—paragraph three of the paper. *Minor premise:* The method of reference [3] was known—list of references. *Conclusion:* The method of reference [3] did not correlate the test data.

The purpose of submitting the paper was to introduce a new correlation method which did agree with the test data and which had the other advantages noted. No further contribution would have been made by discoursing on the shortcomings of other methods with respect to the new data. However, a reasonable view of the facts would show that the method of reference [3] was designed to work for seals of larger clearances, blade spacings, and thicknesses than those tested by the author. Since it is well known that gas flow in very small passages usually has different characteristics than flow through similar large passages, one could not logically expect the accuracy of the correlations of reference [3] to tolerate extrapolation into the small clearance range. The size relationship is shown in Fig. 13.

No further general comments on Mr. Egli's discussion are required, but his outline of tests which might have been more "properly" conducted on the author's apparatus contain some erroneous speculations which will be briefly noted so that other experimenters may be apprised of areas requiring special attention.

Mr. Egli states that seal geometry is kept effectively constant if the ratios (S/h) and (t/h) are kept constant with the values of S, t, and h changing. The author found this not to be true.

Thickness-to-clearance ratio is not a suitable single argument, even at a constant pressure ratio, for defining the contraction coefficient. This is fully evident from Fig. 18 of reference [3] in which Mr. Egli gave contraction coefficients for two blade thicknesses as functions of the clearance-to-thickness ratio. Mr. Egli's figure shows that the curves coincide for the two blades for clearances greater than four times the thickness, but that separate curves are required for the two blades for clearances between one and four times the blade thickness. Mr. Egli himself has therefore indicated the inaccuracy of this particular hypothesis. The author's experiments involved clearances relatively smaller than those shown by Egli thereby making the ratio correlation even less applicable.

The contraction coefficients can be expressed with suitable accuracy only by plotting an individual curve for each particular combination of thickness and clearance, and there is no advantage in the use of the ratio. These individual curves can be plotted against pressure ratio or against Reynolds number as shown in Fig. 8. The reasons for preferring Reynolds number in this instance are discussed in the paper.

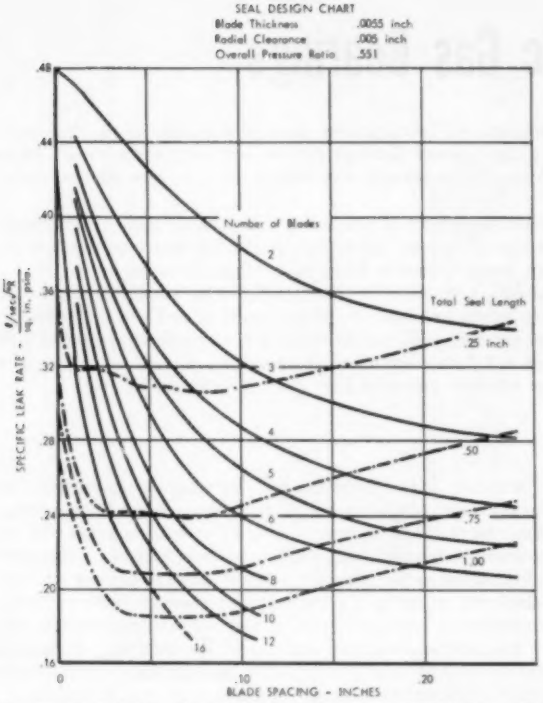


Fig. 14 Seal design chart; blade thickness = 0.0055 in., radial clearance = 0.005 in., pressure ratio = 0.551

As stated in the paper, the velocity carry-over effect is described by the slopes of the experimental curves as in Fig. 7. If this effect can be adequately represented by a multiplier as suggested in Equation (10), the experimental curves of Fig. 7 would have to be straight lines through the origin. The data do not conform to such a straight-line interpretation as may be seen in the figure. Furthermore, if any generalization is to be permitted the experimental correlation curves for given spacing-to-clearance ratios would be required to have a constant slope for varying blade thicknesses and clearances. The data do not conform to this requirement within the standards of accuracy set for these tests.

Finally, Mr. Egli bewails the use of Reynolds number as a correlation criterion solely on the basis of his assumed lack of viscous effects while he also astutely notes that the determination of contraction coefficients in multiblade seals is rather "tricky." Briefly, viscous effects are frequently significant in small channel flow; Reynolds number is not solely a viscosity parameter but expresses the relative importance of inertia and viscous forces; and, as shown in the paper, Reynolds number similarity is the key which permits direct, rational definition of the multiblade contraction coefficient on the basis of single blade test data.

The use of the pressure ratio function is justified by the results, and the relative merit of continuous annular paths compared to bladed seals is discussed with respect to Fig. 14.

Mr. Foster's comments about the effect of blade height show the significance of a variable which was not studied in the author's test program. It was demonstrated that the blade height used was great enough to have no influence on the single blade contraction coefficient determinations, and small enough to preclude bending of the thin blades from high pressure differentials. Fig. 3 is a scale drawing of a typical seal configuration.

Mr. Foster's presentation of their test data for the leak rates of a four-blade seal indicates one method of plotting seal test data. It is interesting to note that these curves of Fig. 11 show the critical pressure ratio of a four-blade seal to be near 0.40, and that this is essentially in agreement with the prediction of Fig. 1, even though the seal has relatively few blades.

The data plot of Fig. 11 is designed to allow ready determinations of the effects of clearance and pressure changes on the operation of a fixed seal design. The method used by the author to present data which could not be correlated by the method of the paper is shown by Fig. 14. This plot is designed to allow selection of optimum seal geometry for a fixed clearance and pressure ratio.

The plot refers to a fixed clearance, fixed blade thickness, and fixed pressure ratio and shows the leak rates for various numbers of blades at a range of interblade spacings. It is possible to superimpose lines of constant total seal length on this plot as shown by the broken lines. These lines then allow a designer to optimize the configuration which will fit into a fixed total length of seal. It is evident that less leakage will result from fitting six or seven equally spaced blades into a 0.50-in. space than would result from using only two or three blades. It is further evident that there is little value in putting more than seven blades into the same space, since no reduction in leak rate would result.

The intersections of the broken lines of constant total seal length with the "zero" spacing axis show the experimentally determined leak rates for uniform annular passages of the clearance and length shown. Thus a uniform annular passage 0.50 in. long and 0.005 in. high would leak 50 per cent more than would six equally spaced blades placed in the same space with the same clearance. As long as the blades are thin and rectangular and the clearances small, blades are preferable to continuous annular passages even if the total seal length is as small as 0.040 in. and the pressure ratio as high as 0.83.

JOHN H. LAUB
Research Group Supervisor,
Jet Propulsion Laboratory,
California Institute of Technology,
Pasadena, Calif.

Hydrostatic Gas Bearings[1]

Orifice-regulated hydrostatic gas bearings offer significant advantages for instrument applications. In particular, gimbal bearings for inertial guidance systems can be designed with negligible starting torque and high stiffness, and for operation at extreme temperatures.

A literature search revealed the lack of convenient and accurate data for the design of hydrostatic gas bearings of various configurations, taking into consideration the effects of compressibility, which cannot be neglected at higher pressures.

Based on Euler's equation, expressions for the significant parameters, i.e., pressure profile, gas-flow rate, gap height, and load-carrying capacity of pad and step bearings, are developed. These parameters yield results which are in excellent agreement with experimental data. The test fixture incorporates pneumatic loading by means of a bellows-suspended piston which is prevented from cocking by an air bearing.

Introduction

INCREASING INTEREST has recently been focused on lubrication of bearings with compressible gases, which has inherent advantages over lubrication with incompressible fluids. Frictional torques and losses of bearings are proportional to the viscosity of the lubricant. The viscosity of gases and vapors is lower, by several orders of magnitude, than that of oils or greases, resulting in a substantial reduction of running friction torques and losses. The starting torque of a well-designed hydrostatic gas bearing is practically zero if coulomb or dry friction is eliminated. Gases are well suited for lubrication at high temperatures at which hydrocarbons will become unstable. Contamination and soiling can be avoided by the use of gases or vapors as lubricants.

Air-lubrication was first suggested by Hirn [1][2] in 1854. During the succeeding half century, it was practically forgotten until, in 1897, Kingsbury published the results of his extensive experiments with an air-lubricated journal bearing [2]. Since then, the state of the art has progressed at an accelerating rate, and the growing interest in this form of lubrication is reflected in the steadily increasing number of publications dealing with the subject. (See Bibliography.)

A survey of the literature revealed much useful information for the design of gas bearings, but also the lack of complete design data and of a rigorous treatment of hydrostatic bearings with restriction of the gas flow by single or multiple orifices. Compressibility effects on the pressure profile within the bearing are very significant as far as the performance characteristics at higher pressures are concerned, namely, the load-carrying capacity, rate of gas consumption, and stiffness of the bearing. A rigorous analysis should, therefore, treat the gas-lubricant as a compressible fluid subject to isothermal expansion.

Characteristics of Circular Pad Bearings

The study was started with an analysis of the characteristics of circular pad bearings. They are individually interesting as thrust bearings but can also be considered the basic modules of more complex bearing configurations, including journal bearings of the type illustrated in Fig. 1.

A pad bearing is schematically shown in Fig. 2. A circular plate of radius R carries a load L and is supported by a compressed-gas pad in a gap. The pad, which has a height h_0, is connected to a gas plenum Pl through an orifice O of diameter D and a cylindrical passage of radius R_c. The gas is supplied to the plenum through an inlet at a volumetric flow rate Q, at the plenum pressure P_{pl}. From the plenum, the gas flows through the orifice O and the passage C to the center of the pad and, from there, radially outward. The flow through the orifice O causes a drop from the plenum pressure P_{pl} to P_0; i.e., the pressure at the center of the pad. This pressure drop, $\Delta P = P_{pl} - P_0$, is a function of the flow rate Q. It serves to regulate the pressure within the gas pad with varying load L and gap height h_0. If the load increases, the gap decreases, and with it the flow rate Q; hence, the pressure drop ΔP becomes smaller and P_0 larger since P_{pl} is constant. If the

[1] This paper presents the results of one phase of research performed at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. DA-04-495-Ord-18, sponsored by the Department of the Army, Ordnance Corps.

[2] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS, and presented at the ASME-ASLE Lubrication Conference, New York, N. Y., October 20-22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, December 12, 1958. Paper No. 59—Lub-1.

Nomenclature

A_1 = area of flow upstream of orifice
A_2 = area of orifice opening
C_D = orifice coefficient
dr = width of infinitesimal circular element of pad
F = force
h_0 = height of compressed-gas pad
I = an integral
L = load on a circular plate
P = pressure at radius r
P_{cr} = critical pressure
P_{pl} = gas plenum pressure; absolute pressure of the gas upstream
P_0 = absolute pressure at inlet condition ($r = R_0$)
P_1 = exit pressure
Pl = gas plenum
p = pressure
Q = volumetric flow rate
R = radius of circular plate
r = radius of infinitesimal circular element of pad
W_{or} = mass flow rate through orifice
W_0 = mass flow rate through gap
$\delta p/\delta x$ = pressure gradient
ΔP = pressure drop
μ = absolute viscosity of gas
ρ_{pl} = density of gas at pressure P_{pl}

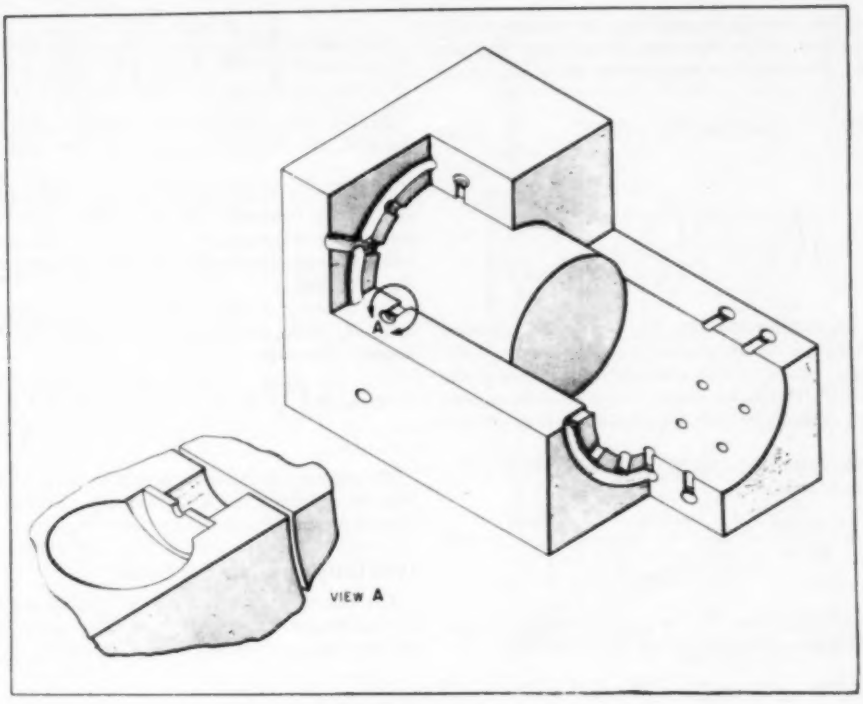


Fig. 1 Schematic of hydrostatic gas bearing

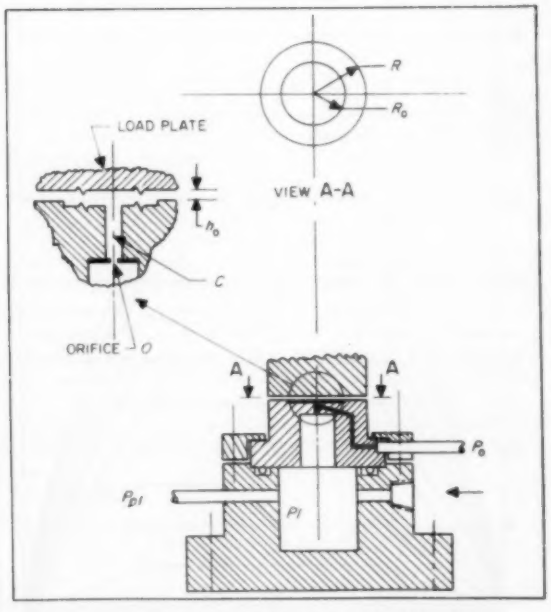


Fig. 2 Diagram of orifice and load-plate of pad bearing

load decreases, the process is reversed. The result is an automatic pressure regulation for varying load.

The load capacity of the pad obviously depends upon the radial pressure distribution in the pad; i.e., the pressure profile $P = f(r)$. It can be calculated from Euler's equation for steady one-dimensional flow with friction, neglecting the body forces of weight and inertia:

$$\frac{\delta p}{\delta x} = -\mu \frac{\delta^2 u}{\delta y^2} \tag{1}$$

in which p denotes pressure; μ, absolute viscosity of gas; $\delta p/\delta x$, pressure gradient; and u, velocity in direction of flow.

If Euler's equation is applied to an infinitesimal circular element of the pad with radius r, width dr, and height h_0, and if isothermal expansion of the gas is assumed, the volume rate of flow and the pressure inside of the pad can be found from the equation

$$\frac{\delta p}{\delta r} = -\frac{12\mu}{2\pi r h_0^3} Q_0 \frac{P_0}{P} \tag{2}$$

where Q_0 is the volume flow rate and P_0 the (absolute) pressure at inlet condition ($r = R_0$), and P is the pressure at radius r. Integration from $r = R_0$ to $r = R$ yields an expression for Q_0

$$Q_0 = \frac{\pi h_0^3}{6\mu \ln \dfrac{R}{R_0}} \left(\frac{P_0^2 - P_1^2}{2P_0} \right) \tag{3}$$

and for P; i.e., the radial pressure distribution in the pad,

$$P = P_0 \left\{ 1 - \frac{\ln \dfrac{r}{R_0}}{\ln \dfrac{R}{R_0}} \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right] \right\}^{1/2} \tag{4}$$

The mass flow rate through the gap $W_0 = \rho_0 Q_0$, or

$$W_0 = \frac{1}{\ln \dfrac{R}{R_0}} \frac{\pi}{12\mu} h_0^3 \rho_0 P_0 \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right] \tag{5}$$

The mass flow rate through the gap must, for reasons of continuity, equal the mass rate of flow through the orifice since the two are in series. The mass flow rate through the orifice can be expressed as

$$W_{or} = C_D \phi A_2 [2\rho_{pl}(P_{pl} - P_0)]^{1/2} \tag{6}$$

where

$$\phi = \left[r^{2/k} \left(\frac{k}{k-1} \right) \left(\frac{1 - r^{\frac{k-1}{k}}}{1 - r} \right) \left(\frac{1 - \left(\frac{A_2}{A_1} \right)^2}{1 - \left(\frac{A_2}{A_1} \right)^2 r^{2/k}} \right) \right]^{1/2} \tag{7}$$

and C_D denotes the orifice coefficient; P_{pl}, the absolute pressure of the gas upstream (i.e., in the plenum); ρ_{pl}, the density of the gas at pressure P_{pl}; A_2, the orifice opening; A_1, the area of the flow upstream, $r = P_0/P_{pl}$; and $k = c_p/c_v$, the ratio of the specific heats of the gas at constant pressure c_p to that at constant volume c_v.

By equating the expressions for W_0 [Equation (5)] and W_{or} [Equation (6)], h_0 is obtained:

$$h_0 = \left\{ \ln \frac{R}{R_0} \frac{12\mu}{\pi} \frac{P_{pl}}{P_0^2 \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right]} \frac{W_{or}}{\rho_{pl}} \right\}^{1/3} \tag{8}$$

Since $\rho_0 = \rho_{pl}(P_0/P_{pl})$, and $W_{or} = \rho_{pl} Q_{or}$, where Q_{or} is the volume rate of flow, the gap height h_0 can also be written

$$h_0 = \left\{ \ln \frac{R}{R_0} \frac{12\mu}{\pi} \frac{P_{pl}}{P_0^2 \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right]} Q_{or} \right\}^{1/3} \tag{9}$$

This expression permits the calculation of h_0 for varying pressures P_0 and a given orifice configuration and plenum pressure P_{pl}.

If a short nozzle or orifice of well-rounded approach is used, the rate of flow increases until the velocity of the downstream jet reaches the velocity of sound in the gas. The critical pressure at which this occurs in an orifice of small diameter in air ($k = 1.4$) is $P_{0cr} = 0.528 P_{pl}$.

For values of P_0 lower than P_{cr}, the flow rate Q remains constant, $Q_{or} = Q_{cr}$, and in the critical range ($0 < P_0 < P_{cr}$) the gap height h_{0cr} becomes

$$h_{0cr} = \left(\frac{12\mu}{\pi} Q_{cr} \right)^{1/3} \left\{ \ln \frac{R}{R_0} \frac{P_{pl}}{P_0^2 \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right]} \right\}^{1/3} \tag{10}$$

The pressure distribution, rate of gas flow, and gap height can thus be rigorously calculated for any value of P_0 between the plenum pressure P_{pl} and exit pressure P_1.

Load-Carrying Capacity of Bearing

The load-carrying capacity of the bearing as a function of P_0 is of considerable interest. The capacity can be calculated from the force balance between the load W and the force F exerted on

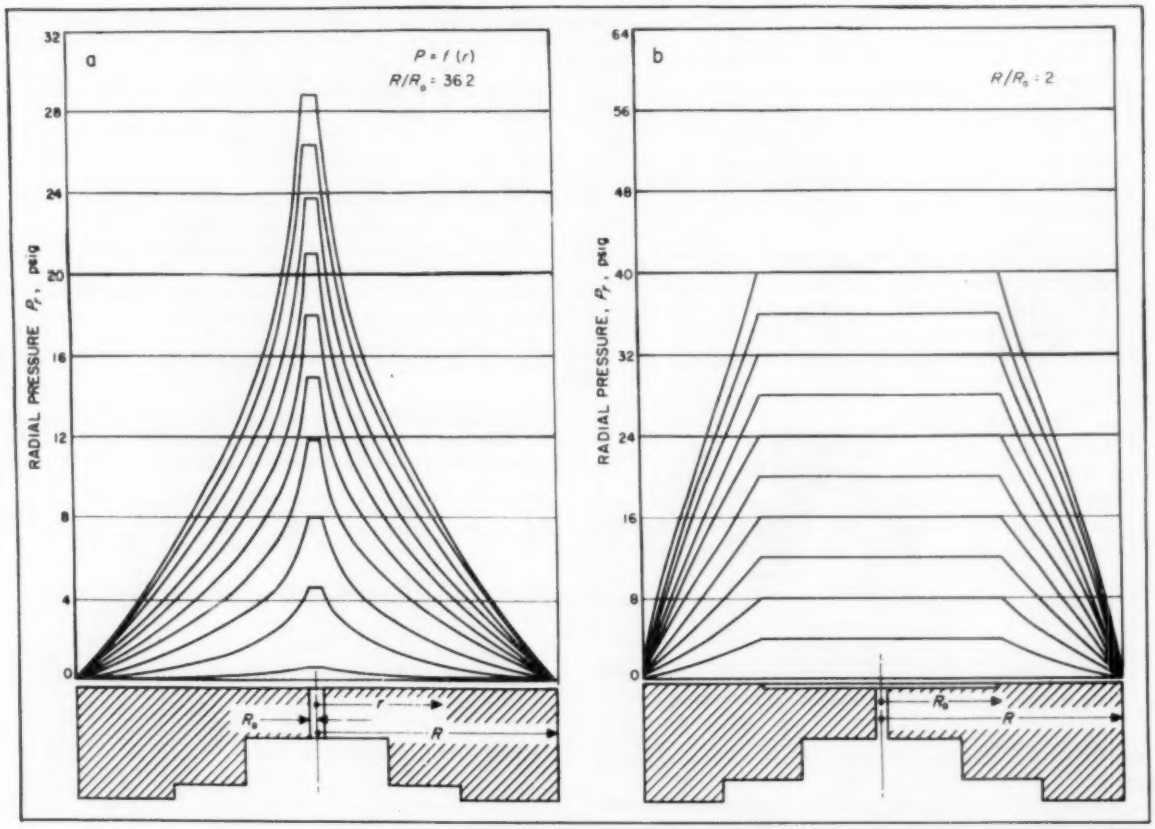


Fig. 3 Pressure profiles

the load plate by the pressure P of the gas. The radial distribution of P between $r = R_0$ and $r = R$ is given by Equation (4). From $r = 0$ to $r = R_0$, the pressure is constant and equal to P_0. Typical pressure profiles for pad bearings with $R = 0.565$ in., and $R_0 = 0.0156$ in. and $R_0 = 0.283$ in., respectively, are shown in Fig. 3.

Since P and P_0 are absolute pressures, the exit pressure P_1 is not zero but must also be considered in F, and the latter can thus be written

$$F = \pi \left(R_0^2 P_0 + 2 \int_{R_0}^{R} P r dr - R^2 P_1 \right)$$

Inserting P from Equation (4):

$$F = \pi P_0 \left\langle R_0^2 + 2 \int_{R_0}^{R} \left\{ 1 - \frac{\ln \dfrac{r}{R_0}}{\ln \dfrac{R}{R_0}} \left[1 - \left(\frac{P_1}{P_0} \right)^2 \right] \right\}^{1/2} r dr - R^2 \frac{P_1}{P_0} \right\rangle \quad (11)$$

In the integral which forms the second term of Equation (11)

$$\frac{1 - \left(\dfrac{P_1}{P_0} \right)^2}{\ln \dfrac{R}{R_0}} = A$$

is constant for a given load, and the integral can therefore be written

$$I = \int_{R_0}^{R} \left\{ 1 - A \ln \frac{r}{R_0} \right\}^{1/2} r dr$$

$$= R_0^2 \int_1^{\frac{R}{R_0}} \left(1 - A \ln \frac{r}{R_0} \right)^{1/2} \frac{r}{R_0} d \frac{r}{R_0}$$

The expression under the square root, or half-power, can be developed into a binomial series

$$1 - \frac{1}{2} A \ln \frac{r}{R_0} - \frac{1}{8} A^2 \ln^2 \frac{r}{R_0} - \frac{3}{48} A^3 \ln^3 \frac{r}{R_0} \cdots$$

which, however, is poorly convergent except for small pressures ($P_1/P_0 \ll 1$). The integration has been carried out to the fifth-power term and to pressures up to $P_0 = 48$ psig. The calculated values show very close agreement with measured values up to pressures of approximately 15 psig; however, for higher pressures, the error increases to about 15 per cent at 48 psig, as would be expected from the poor convergence of the series.

Fortunately, a closed solution for the integral I can be found which yields very good agreement between theory and experiment.

By substituting

$$t = \left(1 - A \ln \frac{r}{R_0} \right)^{1/2}$$

in the expression for I, the integral can be written in the form

$$I = -\frac{2}{A} \int_1^p t^2 e^{(2/A)(1 - t^2)} dt \quad \text{with} \quad p = \frac{P_1}{P_0}$$

$$= \frac{e^{2/A}}{2} \left\{ t e^{-(2/A)t^2} \Big|_1^p - \int_1^p e^{-(2/A)t^2} dt \right\}$$

$$= \frac{e^{2/A}}{2} \left[p e^{-(2/A)p^2} - e^{-2/A} - \sqrt{\frac{A}{2}} \int_{\sqrt{2/A}}^{p\sqrt{2/A}} e^{-(2/A)t^2} \times d\left(t \sqrt{\frac{2}{A}} \right) \right] \sqrt{}$$

$$= \frac{1}{2} \left\{ p e^{2 \ln (R/R_0)} - 1 - e^{2/A} \sqrt{\frac{\pi A}{8}} \left[\mathrm{Erf} \left(p \sqrt{\frac{2}{A}} \right) - \mathrm{Erf} \left(\sqrt{\frac{2}{A}} \right) \right] \right\}$$

The error function Erf is available in tabulated form [3]. With this closed solution for the integral I, the load-carrying capacity F can be written in the form

$$F = \pi P_0 R_0^2 e^{2/A} \sqrt{\frac{\pi A}{8}} \left[\mathrm{Erf} \left(\sqrt{\frac{2}{A}} \right) - \mathrm{Erf} \left(p \sqrt{\frac{2}{A}} \right) \right] \quad (12)$$

or

$$F = A_R P_0 \left(\frac{1}{r'} \right)^2 e^{2/A} \sqrt{\frac{\pi A}{8}} \left[\mathrm{Erf} \left(\sqrt{\frac{2}{A}} \right) - \mathrm{Erf} \left(p \sqrt{\frac{2}{A}} \right) \right] \quad (13)$$

where

$$r' = \frac{R}{R_0}; \quad p = \frac{P_1}{P_0}; \quad A = \frac{1 - p^2}{\ln r'}; \quad A_R = \pi R^2$$

The expression

$$H = P_0 \left(\frac{1}{r'} \right)^2 e^{2/A} \sqrt{\frac{\pi A}{8}} \left[\mathrm{Erf} \left(\sqrt{\frac{2}{A}} \right) - \mathrm{Erf} \left(p \sqrt{\frac{2}{A}} \right) \right] \quad (14)$$

has been plotted for various ratios of R in Fig. 4 as a function of P_0 for pressures P_0 between 8 and 48 psig. From this dimensionless presentation, the load F, which a circular bearing of radius R

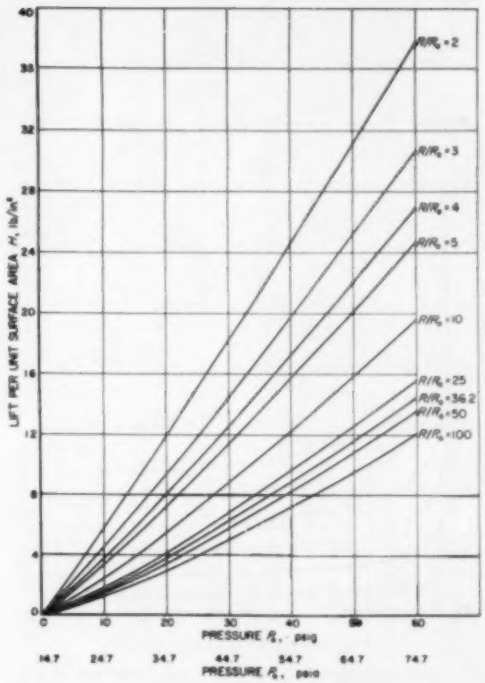


Fig. 4 $H = f(P_0)$, lift per in.^{2} of surface

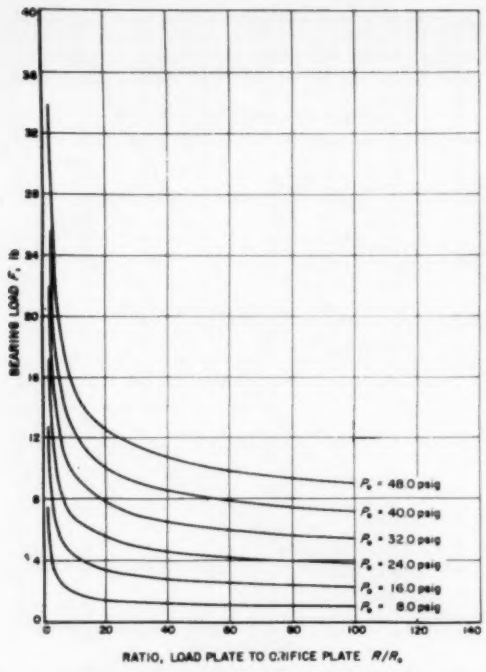


Fig. 5 Load-carrying capacity of step bearing with $\pi R^2 = 1$ in.^{2}

and step radius R_0 can carry at various pressures P_0, can readily be determined.

Fig. 5 shows, as an example, the load-carrying capacity of a circular bearing of one in.^{2} area for a range of R/R_0 from zero to 100 and a pressure range from 8 to 48 psig. It is interesting to note how sharply the maximum-load capability increases at low ratios of R/R_0, corresponding to large step areas. However, this gain in lift is paid for by a decrease in gap height h_0 which, with increasing step radius R_0, soon becomes too small for safe operation of the bearing.

Tests

In order to evaluate the theoretical findings and to compare them with experimental data, a test fixture was built which permits the accurate measurements of the pertinent data of different bearing configurations. The design of this test fixture was deliberately held flexible enough to allow for the testing of pad, step, journal, spherical, and conical bearings, as well as for experiments with various gases and vapors.

Design of Test Fixture. The test fixture consists of an upper load and a lower orifice plate which are separated by an adjustable gap (see Fig. 6). The orifice plate is attached to a plenum chamber which is connected to a regulated and filtered gas supply and mounted on a tool-maker's microscope table. The table can be accurately leveled by means of three leveling screws and aligned with the load plate by means of a pair of x-y lead screws with micrometer adjustment.

The load plate is connected to a piston and prevented from tilting by an air-bearing, which surrounds and guides the piston.

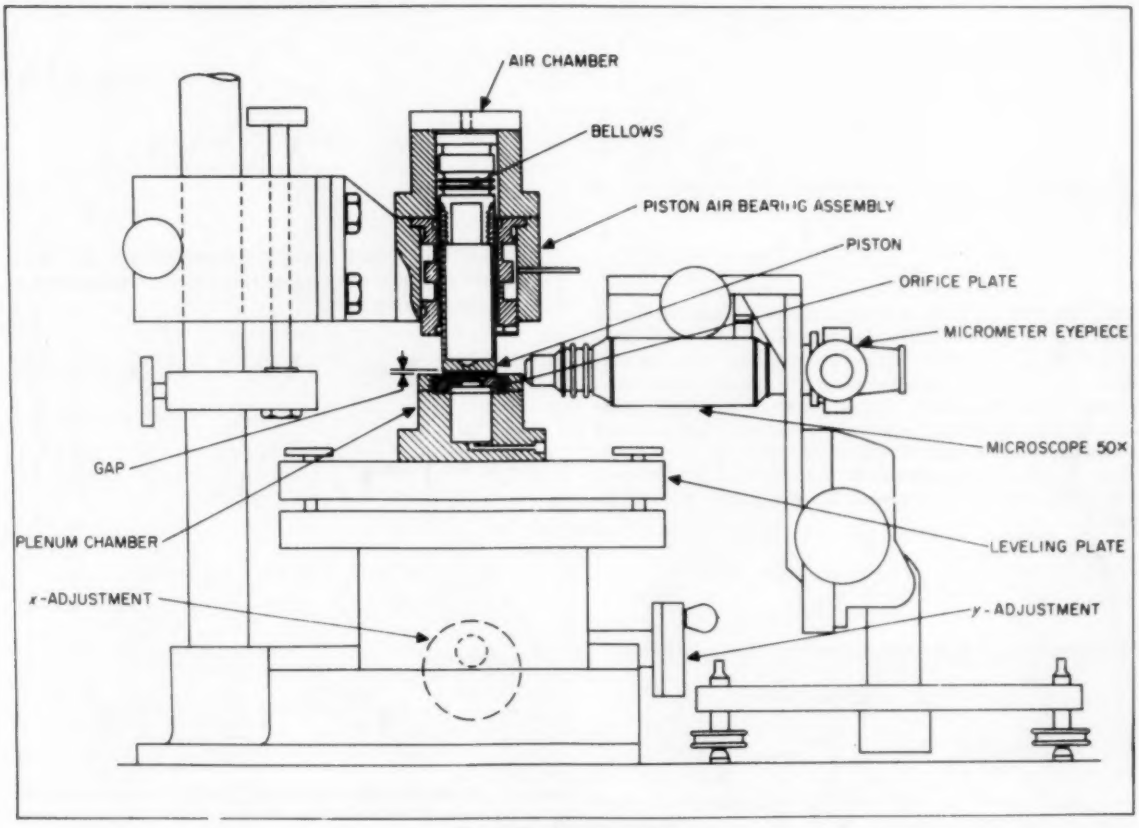


Fig. 6 Gas-bearing test fixture

The piston is connected by means of bellows to an air chamber and can be loaded with varying pressure from a regulated air supply. The load on the load plate can thus be read on a carefully calibrated pressure gage connected to the air chamber. Calibration was accomplished by interposing a strain-gage-type dynamometer between the load plate and the table.

The air chamber, piston, and air bearing are mounted on a support plate, which can be adjusted by a lead screw along a sturdy vertical support column (see Fig. 7).

The orifice plate is fastened to the plenum chamber with a clamp ring and can be easily replaced by other plates with different orifices and contours. Likewise, pistons with different load plates (cylindrical, spherical, or conical) can be connected to the air chamber and their behavior studied under varying conditions of load and pressure.

Fig. 8 illustrates the orifice plate of a pad bearing, and Fig. 9 shows the orifice plate of a step bearing with a 2:1 recess. Spirally arranged tap holes can be seen on the lapped surfaces of both orifice plates; they serve to manometrically measure the radial pressure distribution within the pad and to establish pressure profiles.

As a further example, Fig. 10 shows a semicircular load and orifice plate, with 9 orifice holes spaced 15 deg apart, on which the behavior of journal bearings with multiple orifices can be studied.

Fig. 11 shows the piston-bellows assembly with a circular load plate.

The distance between load plate and orifice plate; i.e., the gap height h_0, can be measured optically by means of a high-power microscope with a micrometer eyepiece. The microscope and a light source for illumination of the gap can be seen in the photo of the complete test fixture (Fig. 12). The same photograph also shows the instrument panel with pressure gages, flow meters, and pressure regulators. This optical method of measuring the gap proved quite satisfactory down to gap heights of approximately 300 μin., but was unsatisfactory below that height. A capacity bridge was, therefore, specially developed (see Fig. 13) with which the capacity of the load-orifice plate cavity could be accurately measured. From this measurement, the gap height was calculated and checked by means of optical measurements with the microscope. This electrical method proved to be highly sensitive, the resolution being approximately one μin.

Experimental Data. The results of a series of measurements of h_0 and Q on three orifice plates of outer radius $R = 0.565$ in. ($\pi R^2 = 1$ in.^{2}) and step ratios of $R/R_0 = 2$, 3, and 36.2, respectively, are plotted against P_0 in Figs. 14 to 17. The plenum pressure in these measurements was varied from 8 to 48 psig in steps of 8 psig. The agreement between experiment (crosses and circles) and theory (solid lines) is very close, considering the limitations of the experimental method.

It is interesting to note in Fig. 17 how the plot of the flow rate Q at a plenum pressure of 32 psig reflects the critical operation of the orifice. The critical pressure P_{cr} for 32 psig, or 46.7 psi absolute, is $P_{cr} = 0.528 \times 46.7 = 24.6$ psi absolute or 9.9 psig. Fig. 17 shows that, for pressures P_0 below this value at a plenum pressure of 32 psig, the flow rate remains constant as the theory demands.

In Fig. 18, the gap height h_0 of the bearing with $R/R_0 = 36.2$ is plotted against the load W. The measurements were again made at various plenum pressures between 8 psig and 48 psig and show how the stiffness of the bearing increases with increasing pressure. If the load, for instance, is 5 lb the spring rate of the bearing operating at a plenum pressure of 32 psig is 6650 lb/in.; it increases to 17,300 lb/in. at 40 psig and to 20,800 lb/in. at 48 psig.

Fig. 19 shows the results of measurements of the flow rate Q with varying load on the bearing with $R/R_0 = 36.2$ at plenum pressures from 8 to 48 psig.

The calculations of the load-carrying capacity [Equations (12) through (14)] were checked by measuring the load which results in closing the gap. Closure loads measured on the three bearings with ratios R/R_0 of 2, 3, and 36.2, respectively, are plotted in Fig. 20 and agree very well with the theoretical prediction represented by the solid lines.

It is interesting to compare the actual load-carrying capacity of a bearing with radius R with the maximum load which the bearing could support if the pressure P did not drop from P_0 to P_1 between R_0 and R but remained constant at a value P_0. This

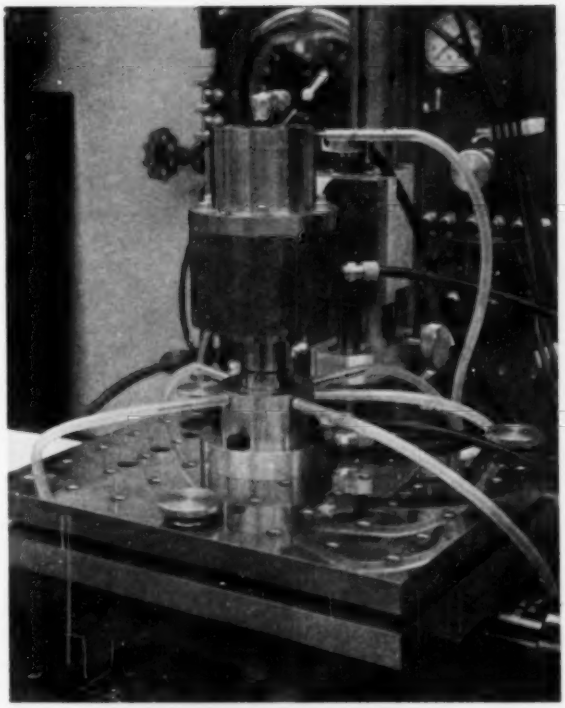


Fig. 7 Test fixture

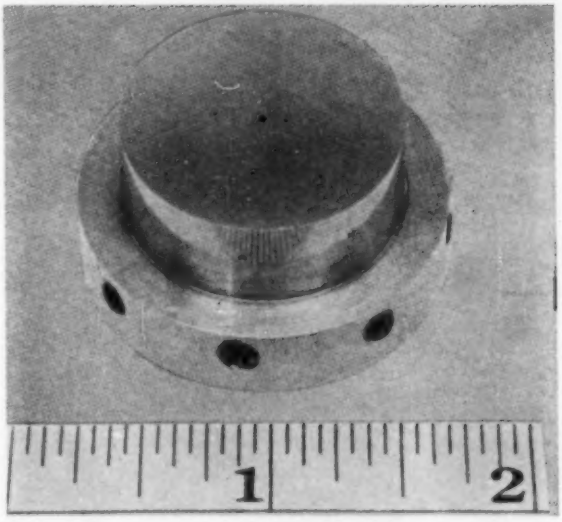


Fig. 8 Pad bearing

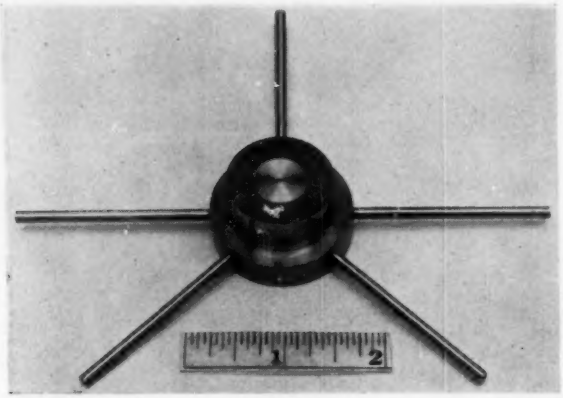


Fig. 9 Step bearing

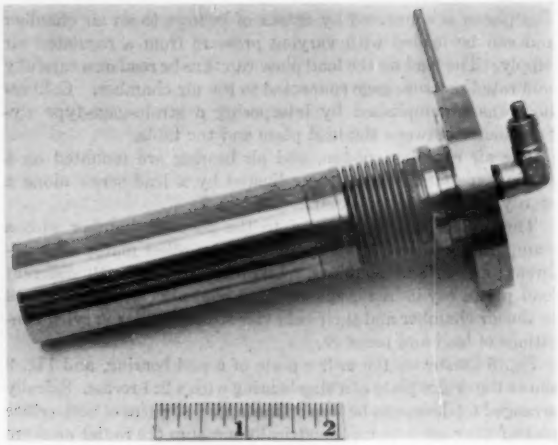
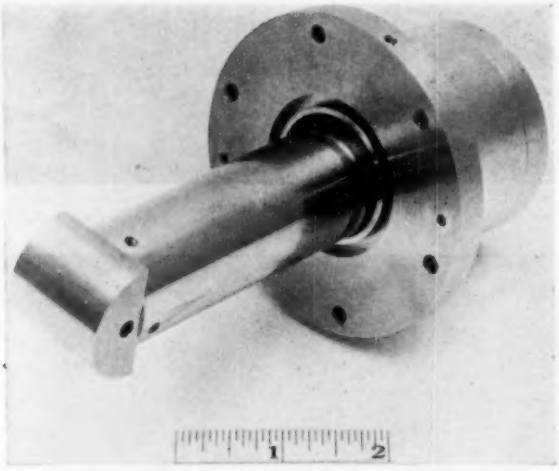


Fig. 11 Piston and bellows



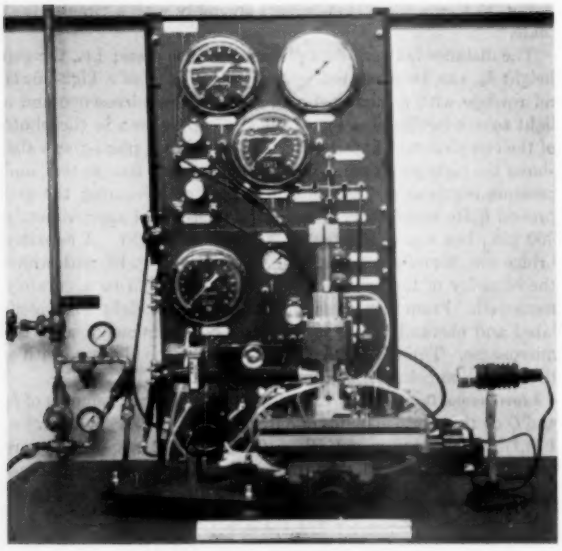


Fig. 12 Gas-bearing test setup

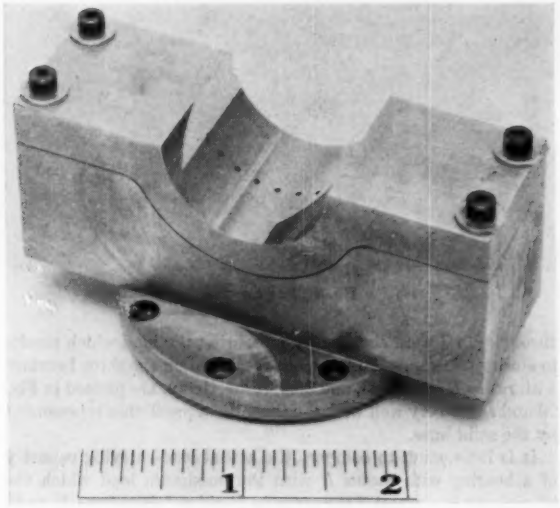


Fig. 10 Semicircular journal bearing

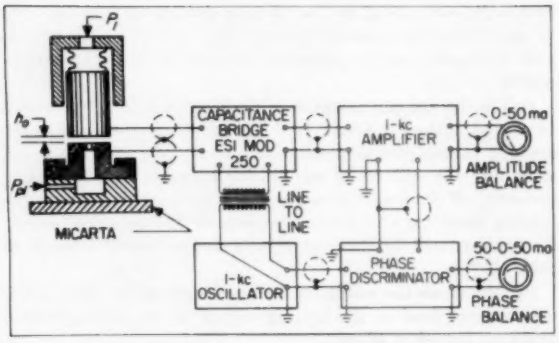


Fig. 13 Capacity bridge for gap measurements

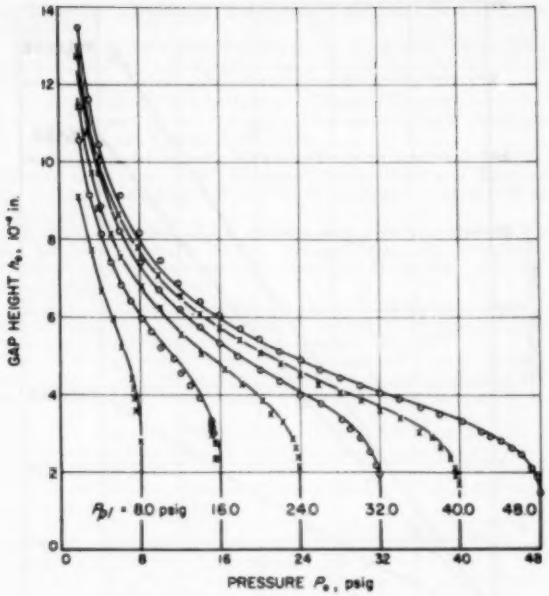


Fig. 14 Gap measurements, pad bearing, $R/R_0 = 36.2$

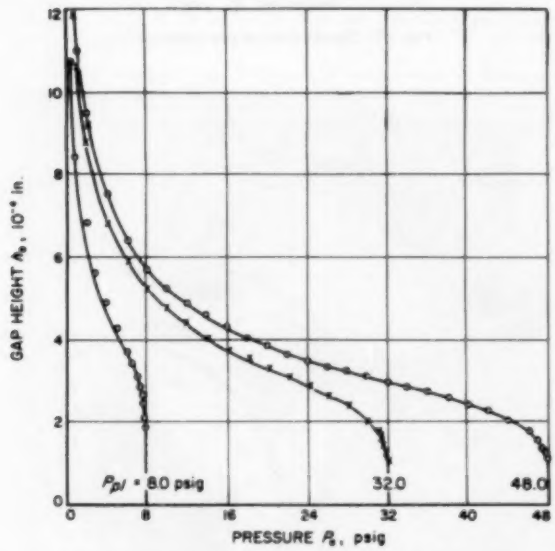


Fig. 15 Gap measurements, step bearing, $R/R_0 = 3$

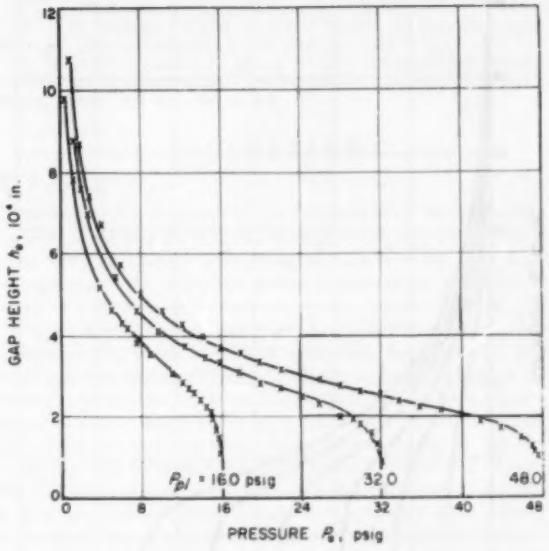


Fig. 16 Gap measurements, step bearing, $R/R_0 = 2$

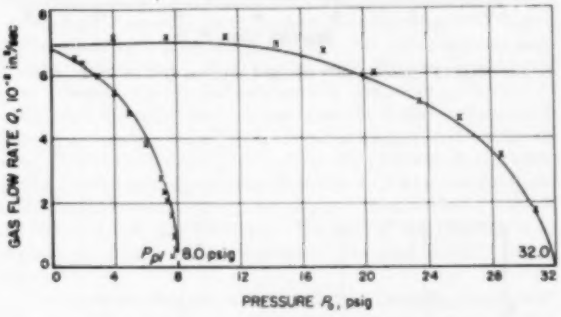


Fig. 17 Pad bearing, $Q = f(P_0)$, rate of gas flow

ratio is sometimes called the "efficiency" of the bearing; it is, of course, a function of R/R_0, but varies little with P_0. The efficiency of the three bearings with $R/R_0 = 2$, 3, and 36.2 is 61.5, 49.5, and 22.0 per cent, respectively, at $P_0 = 40$ psig; and 59.5, 46.5, and 19.0 per cent at $P_0 = 20$ psig.

Conclusions

It can thus be concluded that the theoretical expressions derived for h_0, Q, and F describe the performance of hydrostatic pad and step bearings very accurately within the range of pressures and step radii investigated.

The study of hydrostatic gas bearings is now being extended to journal bearings of the type illustrated in Fig. 21. Results of the theoretical and experimental investigation are to be published.

References

1 G. Hirn, "Sur les Principaux Phenomènes Qui Presentent les Frottements Mediats," Bulletin de la Société Industrielle de Mulhouse, vol. 26, 1854, p. 188.

2 A. Kingsbury, "Experiments With an Air Lubricated Journal," *ASME Journal*, vol. 9, 1897, pp. 267–292.

3 B. O. Pierce, "A Short Table of Integrals," fourth edition, Ginn and Company, Boston, Mass., 1956.

Bibliography

1 D. D. Fuller, "Theory and Practice of Lubrication for Engineers," John Wiley & Sons, Inc., New York, N. Y., 1956.

2 "The Pressurized Fluid Bearing for High Precision Instrument Suspension," Instrumentation Laboratory, M.I.T., Cambridge, Mass., 1948.

3 S. K. Grinnell, "A Study of Pressurized Air Bearing Design; Steady Loading—No Rotation," Department of Mechanical Engineers, M.I.T., Cambridge, Mass., 1954.

4 S. K. Grinnell and H. H. Richardson, "Design Study of a Hydrostatic Gas Bearing With Inherent Orifice Compensation," Dynamic Analysis and Control Laboratory, M.I.T., Cambridge, Mass., 1955.

5 H. H. Richardson, "Behavior of Compensated Hydrostatic

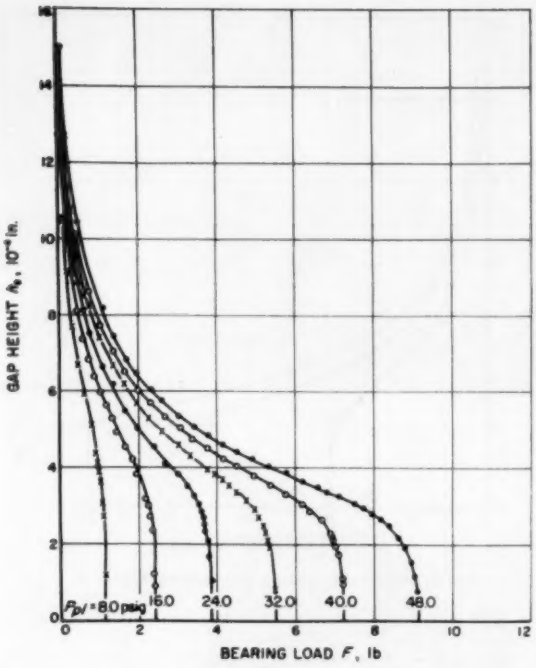


Fig. 18 Bearing stiffness, pad bearing, $R/R_0 = 36.2$

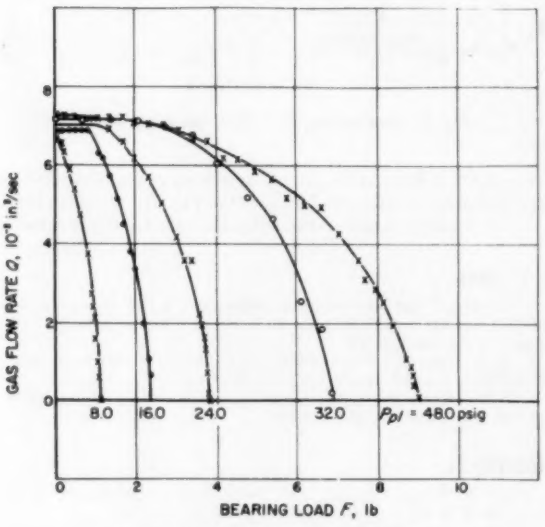


Fig. 19 Rate of gas flow for pad bearing

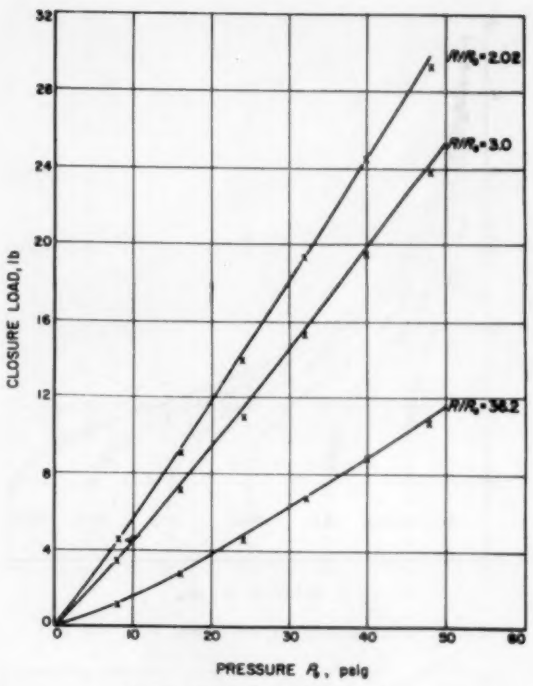


Fig. 20 Closure load of pad bearings

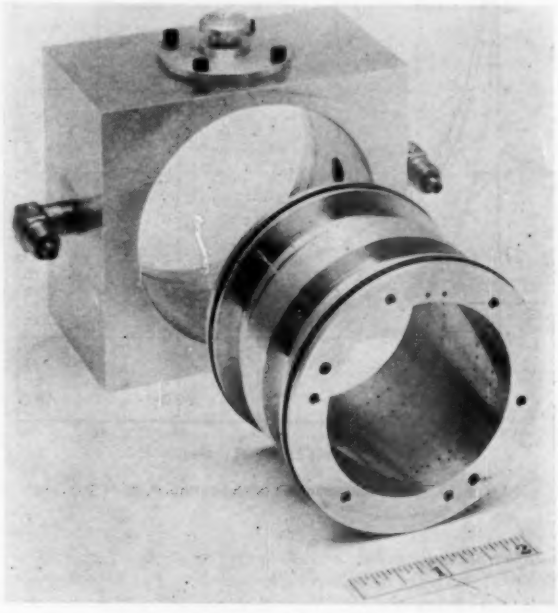


Fig. 21 Hydrostatic journal bearing

Gas Bearings," Mechanical Engineering Department, M.I.T., Cambridge, Mass.

6 A. A. Raimondi and John Boyd, "An Analysis of Orifice Compensated Hydrostatic Journal Bearings," ASME Paper No. 54—Lub-17.

7 R. R. Weber, "The Analysis and Design of Hydrodynamic Gas Bearings," Report AL-699, Aerophysics Laboratory, North American Aviation, Inc., Los Angeles, Calif., 1949.

8 F. Gottwald and R. Vieweg, "Berechnungen and Modellversuche an Wasser—und Luftlagern" (Calculations and Tests of Water—and Air Bearings), *Zeitschrift für angewandte Physik*, vol. 2, 1950, p. 437.

9 I. F. Gottwald, "Computations and Measurements on the Air Bearing," Archive 16/15, Foreign Document Evaluation Branch, Ordnance Research and Development Center, Aberdeen Proving Ground, Md., 1943.

10 I. F. Gottwald, "Proposal for an Air Supported Gyroscope," Foreign Document Evaluation Branch, Ordnance Research and Development Center, Aberdeen Proving Ground, Md., 1943.

11 I. F. Gottwald, "Tests on Air Supported Course Gyros," Archive 16/17, Foreign Document Evaluation Branch, Ordnance Research and Development Center, Aberdeen Proving Ground, Md., 1943.

12 G. F. Midwood, "Tests on an Air Lubricated Thrust Bearing," Aero Technical Memo No. 5, Royal Aircraft Establishment, Farnborough, England, 1947.

13 P. M. Mueller, "Air Lubricated Bearings," *Product Engineering*, vol. 22, no. 8, 1951, pp. 112–115.

14 Sasaki, et al., "On the Effects of a Restrictor Before the Air-Hole of an Air-Lubricated Bearing," *Journal of the Japan Society of Mechanical Engineers*, vol. 20, 1954, pp. 105–108.

15 F. R. Archibald, "A Look at Hydrostatic Thrust Bearings," *Machine Design*, vol. 25, no. 9, 1953, pp. 170–175.

16 T. L. Corey, C. M. Tyler, H. H. Rowand, Jr., and E. M. Kipp, "Behavior of Air in the Hydrostatic Lubrication of Loaded Spherical Bearings," TRANS. ASME, vol. 78, 1956, pp. 893–898.

17 L. Licht, D. D. Fuller, and B. Sternlicht, "Self-Excited Vibrations of an Air-Lubricated Thrust Bearing," TRANS. ASME, vol. 80, 1958, p. 411.

18 P. M. Mueller, "Air Turbine Driven Spindles for Internal Grinding Machines," *Product Engineering*, vol. 23, no. 3, 1952, pp. 160–163.

19 "Bearings for Gyroscope Rotors," *The Engineer*, vol. 194, 1952, p. 651.

20 W. J. Harrison, "The Hydrodynamical Theory of Lubrication With Special Reference to Air as a Lubricant," *Trans. Cambridge Philosophical Society*, vol. 22, 1913, pp. 39–54.

21 W. J. Harrison, "The Hydrodynamical Theory of Lubrication of a Cylindrical Bearing Under Variable Load and of Pivot Bearings," *Trans. Cambridge Philosophical Society*, vol. 22, 1913, p. 373.

22 D. D. Fuller, "Air Bearings—Low Friction," *Lubrication Engineering*, vol. 9, December, 1953, pp. 298–301.

23 J. D. Pigott and E. F. Macks, "Air Bearing Studies at Normal and Elevated Temperatures," *Lubrication Engineering*, vol. 10, February, 1954, pp. 29–33.

24 J. M. Slater, "Gyroscopes for Inertial Navigators," *Mechanical Engineering*, vol. 79, 1957, pp. 832–857.

25 "Piezo-Electric Pick-Up Provides Accurate Vibration Measurement," *Design News*, October 1, 1957, pp. 52–53.

26 R. E. Randall, "Thermodynamic Properties of Air," TR-57-8, Arnold Engineering and Development Center, Gas Dynamics Facility, Department of the Air Force, Washington, D. C., August, 1957.

27 S. K. Grinnell, "Flow of a Compressible Fluid in a Thin Passage," TRANS. ASME, vol. 78, 1956, pp. 765–771.

28 A. H. Shapiro, "The Dynamics and Thermodynamics of Compressible Fluid Flow," The Ronald Press Company, New York, N. Y., vols. I and II.

29 B. Blaschitz, "Development of Design Information for Externally Pressurized Air Bearings," thesis, Mechanical Engineering Department, M.I.T., Cambridge, Mass., 1956.

30 H. F. Brubach, "Some Laboratory Applications of the Low Friction Properties of the Dry Hypodermic Syringe," *Review of Scientific Instruments*, vol. 18, 1947, pp. 363–366.

31 W. Froessel, "Journal Bearings With Hydrodynamic Lubrication," *Stahl und Eisen*, vol. 71, 1951, pp. 125–128.

32 J. W. Beams, "High Rotational Speeds," *Journal of Applied Physics*, vol. 8, 1937, pp. 795–806.

33 G. S. Perkins, P. R. Vogt, and R. R. Weber, "Double-Ended Journal Air Bearing," U. S. Patent No. 2,597,371, 1952.

34 E. R. Penick, "Bearing," U. S. Patent No. 1,906,715, 1933.

35 L. F. Carter, "Air Supported Gyro," U. S. Patent No. 2,000,896, 1937.

36 W. J. Abbot, Jr., "Air-Lubricated Bearings," U. S. Patent No. 1,337,742, 1920.

37 M. Wildmann, discussion on paper by J. S. Ausman, p. 1224.

38 J. S. Ausman, "Fluid Dynamic Theory of Gas-Lubricated Bearings," TRANS. ASME, vol. 79, 1957, p. 1218.

39 B. Sternlicht and R. C. Elwell, "Theoretical and Experimental Analysis of Hydrodynamic Gas-Lubricated Journal Bearings," TRANS. ASME, vol. 80, 1958, p. 865.

DISCUSSION

W. A. Gross[3]

The author has presented a clear discussion of the externally pressurized circular-plate thrust bearing which is orifice compensated. He has made a unique contribution in conducting careful experiments to verify the known equations for this type of flow.

A word about the limitations involved is in order. Inasmuch as fluid inertia terms have been neglected, the equations lose validity in the region of the inlet hole as the inlet pressure is increased or as the film thickness is increased. In addition, the existence of sonic velocity at the inlet radius R_0, followed by supersonic velocities, and eventually shock waves was first recognized by Deuker and Wojtech,[4] and has since been analyzed by others. Since the expansion occurs so rapidly in the inlet section, the flow there is more nearly adiabatic whereas throughout the remainder of the space between the two surfaces it approaches isothermal.

The effect of fluid inertia is of great importance for externally pressurized bearings, not only because of the modifications in the equation of fluid motion, but also because it can induce turbulence. For example, transition from laminar flow can occur with Reynolds numbers as low as 550 based upon the film thickness and average film velocity. The applicability of these limitations must be assessed before the equations in this paper are used.

Although the paper concentrates upon the orifice compensated bearing, as illustrated in Fig. 1, the pressure and load results as given in Section 3 depend only upon the preasure at the inlet section. For example, the left half of Fig. 3 shows a configuration which is compensated only by the change of film thickness around the periphery of the inlet hole. Because of the absence of a trapped air volume this configuration has good stability characteristics.

This discusser has conducted tests on externally pressurized flat-plate thrust bearings by fixing the plate which contains the inlet hole and moving a plate containing a one-mil diameter pressure sensing hole so that pressure at any radial position may be measured. The small pressure sensing hole is necessary in order to detect the large pressure gradients which can occur in the vicinity of the inlet hole. The pressures can fall below ambient pressure in a distance of a few mils. Too much modification of the flow, as well as effectively averaging the pressure, occurs if the pressure sensing holes are large.

Author's Closure

I appreciate the comments of Dr. W. A. Gross.

As far as the limitations of the analytical treatment chosen in the paper are concerned, it is the opinion of the author that it is not only valid as applied to bearings which are of interest to instrumentation but also to many other practical applications of externally pressurized gas bearings.

The assumptions made in the analysis are that the flow is in the laminar regime throughout the film, that inertia forces are negligibly small compared with viscous forces and that the gas expands isothermally. The very good correlation of the experimental and theoretical data is proof that these assumptions are fully justi-

[3] Member of Research Staff, Applied Mechanics, IBM Research Laboratory, San Jose, Calif. Mem. ASME.

[4] "Ecoulement Radial d'un Fluide Visqueux Entre Deux Disques Très Rapprochés; Théorie du Palier a l'Air," *Revue Générale de l'Hydraulique*, vol. 17, pp. 227–234, 285–294.

fied in the case of orifice regulated hydrostatic bearings used as pivot axis supports in inertial and similar instruments which operate with clearances of approximately 1.5 mil and less, orifice diameters of about 4 mil and supply pressures up to 50 psig.

The effects of fluid inertia, turbulent flow, and shock waves are undoubtedly of considerable academic interest; however, as far as practical applications are concerned, the author cannot agree that it is of great importance. As Comolet[5] and others have shown, the load carrying capacity approaches zero or can even become negative when critical values of the Reynolds and Mach numbers are exceeded; the bearing, therefore, loses its usefulness completely. Clearances of 4 to 20 mil and more, which have been used by several authors to demonstrate these abnormal effects, at high supply pressures and excessively large flow rates, are hardly of interest to the designer of bearings for practical applications who is concerned with satisfactory load carrying capacity, maximum possible stiffness, and economy of gas consumption.

[5] Raymond Comolet, "Ecoulement d'un Fluid Entre Deux Plans Parallels. Contribution a l'Etude de Buttees d'Air," Service de Documentation et d'Information de l'Aeronautique, Paris, France.

Effect of Hardness, Surface Finish, and Grain Size on Rolling-Contact Fatigue Life of M-50 Bearing Steel

R. A. BAUGHMAN
Manager-Materials Applications Unit, General Electric Company, Flight Propulsion Laboratory Department, Cincinnati, Ohio

The effect of hardness, surface finish, and grain size upon the compressive rolling-contact fatigue strength of M-50 bearing steel has been studied. Considerable testing on the RC Rig and statistical-treatment methods have been included. A mathematical expression relating these variables to life expectancy is presented and the optimization of these variables is discussed. It is shown that bearing fatigue of M-50 increases by increasing hardness, decreasing surface, and increasing grain size. The optimum life identified occurs at R$_C$64 hardness, 1.5 rms surface finish, and a grain size of ASTM 2.

SPECIFICATIONS set for antifriction-bearing procurement, including hardness, grain size, and surface finish are based primarily on results obtained through experience on full-scale bearing tests. This technique is satisfactory and has resulted in excellent progress in bearing development. It is, however, fallacious to assume that bearing tests are the ideal manner to study isolated variables because of the numerous other variables which are part of the manufacturing cycle and which may affect life. The difficulty of isolating one particular variable for the purpose of evaluation by bearing tests is overwhelming. The development of the RC Rig[1] has made possible the isolation and study of *one* variable at a time. In order to determine the significance of these variables as related to bearing life for design and manufacturing purposes, the primary and secondary effects of these variables upon fatigue life must be separated and analyzed. This is necessary in order to establish optimum conditions for these variables. Stress level is another variable included to introduce some flexibility to the applicability of the results.

In order to separate the effects of the variables upon fatigue life, a statistical approach is used. A complete description of the statistical techniques employed in designing this statistical experiment and in analyzing the statistical results is beyond the scope of this paper. It is sufficient to note, however, that the several variables affecting life are separated by various regression methods and by advanced methods of order statistics.

The material used for this investigation was a Vanadium Alloy's air-melted heat of M-50, heat no. 52006. A series of heat-treat studies was conducted in order to obtain the desired range of hardness and grain size. Austenitizing was accomplished at 2050 F with time increasing for increased grain size. Double-tempering treatments were always used with temperature increased to decrease hardness. Hardness was checked by macro and microhardness tests, grain size by metallographic examinations. To obtain the surface-finish range desired, different grades of grinding wheels were used on the centerless grinder.[2] Metallographic polishing techniques were employed on extremely fine surface-finish levels (1.5 rms). Surface finish was checked via standard profilometer instruments.

To eliminate the effect of dissimilar variables in the RC Rig rollers, the rollers were mated identically with each set of test bars. That is, at any given condition the test bars and rollers had identical grain size, hardness, and surface finish.

Mil-L-7808 lubricant at 20 drops per minute (equivalent to flooding) was used throughout the test program. All tests were run at room temperature.

Experimental Test Results

The results of experimental fatigue tests are given in Table 1. Four tests were conducted at each combination of grain size, hardness, surface finish, and stress level. As expected, great variations in fatigue life were exhibited in spite of the fact that four repetitive tests were run under identical conditions. This variation in life (scatter) is due to the property that any finite collection of test results for a fixed set of conditions belongs to some definite population, defined by the manner in which the fatigue lives of the individual members are distributed within the population. An estimate of this population is obtained by constructing a graphical picture of the data commonly plotted in Weibull[3] form which is a plot of the per cent of tests failed as the ordinate versus the cycles to failure as the abscissa. This distribution is linear when plotted on Weibull distribution paper. Values obtained from such plots are given in Table 2.

The use of only four tests for each condition of grain size, hardness, and surface finish is admittedly a bare minimum. Its justification is based upon reproducibility studies, previously published[1] on the RC Rig method for rolling-contact testing. The simplicity of the testing method tends to reduce the scatter and therefore the number of tests necessary to establish the position and slope of the Weibull graph. It has been observed that slope and position of the values studied move only slightly when based on more than four test points.

In order to compare the test results on an equal basis, these results are evaluated and compared at the same stress level. This was accomplished by interpolation of B_{10} and B_{50} lives of the repetitive tests between stress levels, assuming a linear variation between stresses. The lives at 700,000 psi are tabulated in Table 3 according to decreasing B_{10} life, and Table 4 according to de-

[1] R. A. Baughman, "Experimental Laboratory Studies of Bearing Fatigue," ASME Technical Digest, *Mechanical Engineering*, March, 1959, p. 94.

[2] The RC Rig utilizes cylindrical specimens 3 in. long and $^3/_8$ in. diameter.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 18, 1959. ASME Paper No. 59—Lub-11.

[3] W. Weibull, "A Statistical Distribution Function of Wide Applicability," *Journal of Applied Mechanics*, vol. 18, TRANS. ASME, vol. 73, 1951, pp. 293–297.

Table 1 Experimental test results

Hardness (Rc)	Grain Size (ASTM)	Surface Finish (r.m.s.)	Maximum Hertz Stress (10^{-3}psi)	Cycles to Failure x 10^6 #1	#2	#3	#4
64	7	9.35	700	1.121	3.190	5.620	10.429
62 1/4	10.4	3.75	665	1.359	3.409	7.239	12.56
62 1/4	3.4	3.75	665	6.248	12.288	19.04	22.69
62 1/4	10.4	14.75	665	1.001	2.968	6.771	13.144
62 1/4	3.4	14.75	665	2.446	3.469	4.89	11.723
62 1/4	10.4	3.75	732	.991	1.43	3.26	4.95
62 1/4	3.4	3.75	732	3.487	4.199	6.109	10.076
62 1/4	10.4	14.75	732	.961	3.165	7.039	11.24
62 1/4	3.4	14.75	732	.926	2.46	5.82	8.11
58	12	9.75	700	1.011	3.240	4.690	5.875
58	2	9.75	700	.989	2.004	3.960	6.105
58	7	9.75	700	1.74	2.87	4.89	10.29
58	7	1.5	700	1.902	3.284	4.229	5.149
58	7	17	700	2.436	5.356	6.195	7.520
58	7	9.75	645	2.166	3.607	5.07	8.060
58	7	9.75	745	.897	1.002	1.740	2.69
54	10.4	3.75	665	.656	1.309	2.69	2.79
54	3.4	3.75	665	.697	1.11	1.94	2.69
54	10.4	14.75	665	.487	.697	1.80	2.03
54	3.4	14.75	665	.412	1.032	1.724	2.181
54	10.4	3.75	732	.248	.629	1.209	2.667
54	3.4	3.75	732	.379	.862	1.169	1.665
54	10.4	14.75	732	.246	.296	.315	.712
54	3.4	14.75	732	.361	.605	.744	1.435
52	7	9.75	645	.171	.329	.487	1.002

Table 2 Tabulation of B_{10} and B_{50} and the Wiebull slope of R_c Rig fatigue results on M-50 at various grain sizes, hardnesses, and surface finishes

Hardness	Grain Size	Surface Finish	Stress (1)	B_{10} Life	B_{50} Life	Wiebull Slope
64	7.0	9.35	700	.7	4.3	1.11
62 1/4	10.4	3.75	665	.64	3.9	1.12
62 1/4	3.4	3.75	665	4.8	10.3	2.00
62 1/4	10.4	14.75	665	.5	4.4	0.91
62 1/4	3.4	14.75	665	1.4	5.2	1.55
62 1/4	10.4	3.75	732	.56	2.3	1.50
62 1/4	3.4	3.75	732	2.1	6.1	2.10
62 1/4	10.4	14.75	732	.56	4.4	0.97
62 1/4	3.4	14.75	732	.52	3.3	1.09
58	12.0	9.75	700	.78	3.3	1.35
58	2.0	9.75	700	.60	2.9	1.30
58	7.0	9.75	700	1.0	4.3	1.34
58	7.0	1.5	700	1.6	3.7	2.30
58	7.0	17.0	700	2.1	5.2	2.19
58	7.0	9.75	645	1.5	4.4	1.79
58	7.0	9.75	745	.55	1.5	2.21
54	10.4	3.75	665	.50	1.7	1.71
54	3.4	3.75	665	.47	1.6	1.73
54	10.4	14.75	665	.32	1.30	1.79
54	3.4	14.75	665	.35	1.35	1.50
54	10.4	3.75	732	.14	1.00	1.05
54	3.4	3.75	732	.25	.94	1.53
54	10.4	14.75	732	.16	.38	2.07
54	3.4	14.75	732	.25	.74	1.85
52	7.0	9.75	645	.17	.48	1.87

(1) Max. Hertz stress (10-3 psi)

Table 3 Tabulated test results of B_{10} life in descending order

No.	Hardness (Rc)	Grain Size (ASTM)	Surface Finish (r.m.s.)	Maximum Hertz Stress(10^{-3}Psi)*	B_{10} Life** (Cycles)
1	62 1/4	3.4	3.75	700	3.4×10^6
2	58	7	17	700	2.1×10^6
3	58	7	1.50	700	1.6×10^6
4	58	7	9.75	700	1.0×10^6
5	62 1/4	3.4	14.75	700	$.95 \times 10^6$
6	58	12	9.75	700	$.78 \times 10^6$
7	64	7	9.75	700	$.70 \times 10^6$
8	62 1/4	10.4	3.75	700	$.66 \times 10^6$
9	58	2	9.75	700	$.60 \times 10^6$
10	62 1/4	10.4	14.75	700	$.53 \times 10^6$
11	54	3.4	3.75	700	$.33 \times 10^6$
12	54	10.75	3.75	700	$.31 \times 10^6$
13	54	3.4	14.75	700	$.28 \times 10^6$
14	54	10.4	14.75	700	$.25 \times 10^6$
15	52	7	9.75	700	$.17 \times 10^6$

* All data reduced to 700,000 psi from test data.

** Test speed = 30,000 cycles/minute.

Table 4 Tabulated test results of B_{50} life in descending order

No.	Hardness (Rc)	Grain Size (ASTM)	Surface Finish (r.m.s.)	Maximum Hertz Stress(10^{-3}Psi)*	B_{50} Life** (Cycles)
1	62 1/4	3.4	3.75	700	10×10^6
2	58	7	17	700	5.1×10^6
3	62 1/4	10.4	14.75	700	4.8×10^6
4	62 1/4	10.4	3.75	700	4.1×10^6
5	62 1/4	3.4	14.75	700	4.3×10^6
6	64	7	9.75	700	4.3×10^6
7	58	7	1.5	700	3.6×10^6
8	58	12	9.75	700	3.3×10^6
9	58	2	9.75	700	2.8×10^6
10	58	7	9.75	700	2.8×10^6
11	54	3.4	3.75	700	1.2×10^6
12	54	10.4	3.75	700	1.3×10^6
13	54	3.4	14.75	700	1.0×10^6
14	54	10.4	14.75	700	$.80 \times 10^6$
15	52	7	9.75	700	$.48 \times 10^6$

* All data reduced to 700,000 psi from test data.

** Test speed = 30,000 cycles/minute.

creasing B_{50} life. These results indicate that B_{50} life is predominantly dependent upon hardness while B_{10} life is dependent upon the interaction of hardness, grain size, and surface finish.

General Fatigue-Life Equation

The general equation relating hardness, surface finish, and grain size to fatigue life as obtained from a statistical analysis of the test data is of the following form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \ldots + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^3 + \beta_{44} x_4^2 + \beta_{12} x_1 x_2 \ldots + \beta_{13} x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 + \beta_{24} x_4 \ldots + \beta_{34} x_3 x_4$$

where

Y = same measure of fatigue life
X_1 = a function of hardness
X_2 = a function of grain size
X_3 = a function of surface finish
X_4 = a function of maximum Hertzian stress
β = general coefficient to be determined by statistical methods applied to test data

Table 5 Experiment to determine the effect of hardness, grain size, surface finish, and stress on life of bearing materials

Summary of Results

Variables Z_1 = Hardness, Rc; Z_2 = Grain Size, ASTM; Z_3 = Surface Finish, RMS; Z_4 = Stress PSI, Max. Hertz x 10^3

Coefficients of the General Equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{44} X_4^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{14} X_1 X_4 + \beta_{23} X_2 X_3 + \beta_{24} X_2 X_4 + \beta_{34} X_3 X_4$

Based Upon

	Y_1 Arithmetic Mean	Y_2 Mean Log of Values	Y_3 M = Shape Parameter For Weibull Distribution	Y_4 Θ = Scale Parameter For Weibull Distribution	Y_5 10% Life Based on Order Statistics	Z_3 M = Shape Parameter Eliminating Z_3, Z_4
β_0	3786878.8	15.083532	2.3294455	4269427.7	1185578.4	2.6568947
β_1	2212030.3	.72147809	- .13181985	2457279.9	959151.9	- .13273504
β_2	- 468309.3	- .10982447	- .14732217	- 562745.0	- 331716.4	- .14911162
β_3	- 361559.3	- .09883920	- .07671921	- 426937.9	- 245798.9	
β_4	- 1062018.1	- .27899127	- .01760625	- 1189920.1	- 372022.7	
β_{11}	- 329260.1	- .38992290	- .35422400	- 394921.2	- 358933.4	- .34247003
β_{22}	- 10010.3	- .04740334	- .25056545	- 19012.3	- 124135.0	- .25952200
β_{33}	371253.5	.09281592	.35110697	399306.2	341980.6	
β_{44}	- 10873.9	- .12882581	- .02765744	- 6529.1	027343.0	
β_{12}	- 538209.0	- .06651590	- .06876103	- 639619.0	- 366082.2	- .06908010
β_{13}	- 343922.9	.02918470	- .15323202	- 411832.6	- 307471.2	
β_{14}	- 504700.5	.03290228	.06073631	- 556773.7	- 423566.7	
β_{23}	671914.7	.07246280	.11490886	750158.4	332411.8	
β_{24}	335794.6	- .00903349	.01369279	398184.3	211783.7	
β_{34}	576704.3	.03870780	.03517484	642937.9	216832.5	

Identification of Variables

X_1	Hardness - 58.08000 / 3.711307	Same	Same	Same	Same	Same
X_2	Grain Size - 6.936000 / 3.137245	Same	Same	Same	Same	Same
X_3	Surface Finish - 9.390000 / 4.920915	Same	Same	Same	Same	Same
X_4	Stress - 696.440000 / 32.093770	Same	Same	Same	Same	Same

Analysis of Variance

MS Due to Regression	1353382928. x 10^4	1.21756429	.45666929	17.005846&2. x 10^4	2025601928. x 10^3	.68829200
MS Error	388869400. x 10^4	.13502000	.19501600	491042500. x 10^4	90914&500. x 10^3	.25&2&74
F Ratio	3.4803	8.9645	2.3321	3.4632	2.2280	2.663&
Significance Level	5%	1%	10%	5%	Almost 10%	10%

As can be seen from this equation, both the independent and the interaction effects of hardness, grain size, and surface finish are considered. This equation also includes the maximum Hertz stress as a variable.

Results of Statistical Analysis (B_{10}-Life Equation)

The results of the statistical approach are tabulated in Table 5. As can be seen from the table, the various coefficients of the general life equation are evaluated for different measures of fatigue life. One commonly used measure is the B_{10} fatigue life. The B_{10} life of a collection of tests is the life up to which 10 per cent of tests in a population will have failed. This life is of primary interest in that we are generally more interested in being able to predict the occurrence of early failures rather than the magnitude of the mean life. While it is not possible to predict the B_{10} life from a given number of failures with the same degree of confidence as that obtained for the mean life, the statistical results of this analysis are at a significant level of almost 10 per cent.

Representative graphical presentations of the results of the B_{10}-life equation are shown in Fig. 1.

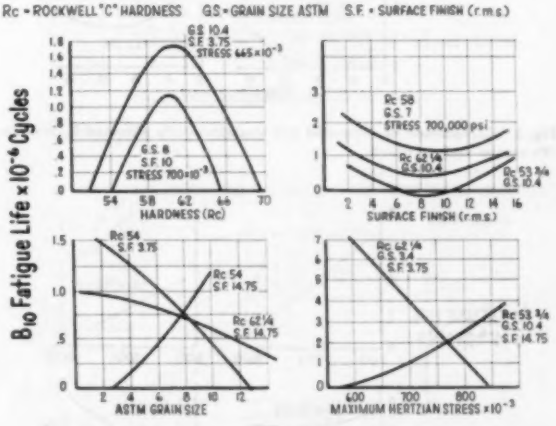

Rc = ROCKWELL "C" HARDNESS
GS = GRAIN SIZE ASTM
S.F. = SURFACE FINISH (rms)
B10 Fatigue Life x 10^-6 Cycles
HARDNESS (Rc)
SURFACE FINISH (rms)
ASTM GRAIN SIZE
MAXIMUM HERTZIAN STRESS x10^-3

Fig. 1 Graphical representation of B_{10} fatigue life equation

Discussion of Results Predicted by the B_{10}-Life Equation

The effects of hardness, stress, surface finish, and grain size upon the B_{10} fatigue life of M-50 as expressed by the B_{10}-life equation are shown in Fig. 1. As can be seen from the curve, this mathematical expression predicts lives which are negative for various combinations of the test variables. This equation also predicts an increase in life as the stress level increases for

other combinations of these variables. These peculiar results may be attributed to the fact that the initial general equation is of a quadratic form for each variable. In effect, this assumes that each variable produces a life which passes through either a minimum or a maximum as the variable is monitorically increased.

Graphical comparisons of the test results and the results obtained from the empirical B_{10}-fatigue-life equation are made in Figs. 2, 3, and 4.

Fig. 2 shows that a close correlation exists between the mathematically predicted effects of grain size upon the B_{10} life and the effects of grain size obtained experimentally.

The effect of surface finish upon life, however, is not clearly defined. Although Fig. 3 shows a good correlation between the mathematically predicted and experimentally obtained lives, only Fig. 3(*d*) substantiates the result that a minimum life occurs at a surface finish of about 10 rms. It is felt that these predicted minimum lives are greatly influenced by a set of high lives obtained at a surface finish of 17 rms (see Table 3). Since only one set of test data is obtained at this relatively coarse surface finish, any conclusions based predominantly upon this data must be questioned.

From Fig. 4, we find that the equation relating stress to life is not valid for certain combinations of grain size, hardness, and surface finish. In view of these shortcomings, the discussion of the results must be restricted within certain limiting values of each variable.

The discussion of results which follow are valid for hardnesses greater than 54 R_c, grain size larger than ASTM 10, and surface finishes better than 10 rms. In order to simplify the B_{10} equation, the stress variable is eliminated by evaluating all lives at a modified Hertz stress of approximately 700,000 psi. The resulting equation is called the Modified-B_{10}-Fatigue-Life equation since it eliminates all of the stress-dependent terms.

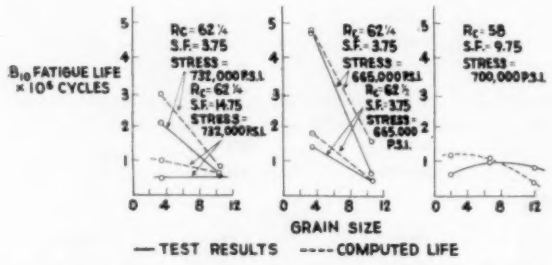


Fig. 2 Comparison of computed and experimentally obtained B_{10} fatigue life versus grain size

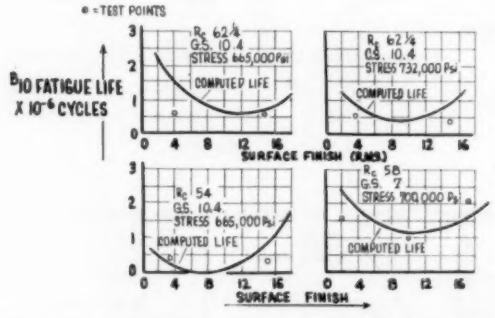

o = TEST POINTS
B_{10} FATIGUE LIFE X 10^{-6} CYCLES
Rc 62¼ G.S. 10.4 STRESS 665,000 Psi COMPUTED LIFE
Rc 62¼ G.S. 10.4 STRESS 732,000 Psi COMPUTED LIFE
Rc 54 G.S. 10.4 STRESS 665,000 Psi COMPUTED LIFE
Rc 58 G.S. 7 STRESS 700,000 Psi COMPUTED LIFE
SURFACE FINISH (RMS)
SURFACE FINISH

Fig. 3 Comparison of computed and experimentally obtained B_{10} fatigue life versus surface finish

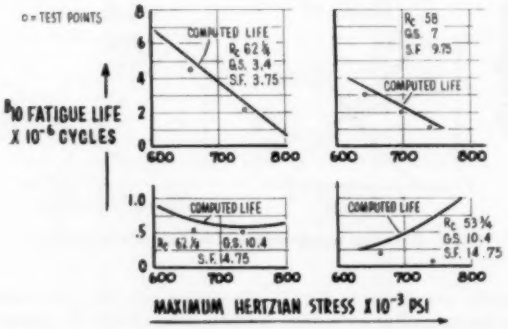

o = TEST POINTS
B_{10} FATIGUE LIFE X 10^{-6} CYCLES
COMPUTED LIFE
Rc 62¼ G.S. 3.4 S.F. 3.75
Rc 58 G.S. 7 S.F 9.75
Rc 62¼ G.S. 10.4 S.F. 14.75
Rc 53¼ G.S. 10.4 S.F. 14.75
MAXIMUM HERTZIAN STRESS X 10^{-3} PSI

Fig. 4 Comparison of computed and experimentally obtained B_{10} fatigue life versus maximum Hertzian stress

The Modified-B_{10}-Fatigue-Life Equation

In order to obtain our modified equation, the general life equation is rewritten as follows:

$$\text{Life} = \beta_0 + x_1[\beta_1 + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3] \ldots + x_2[\beta_2 + \beta_{22}x_2 + \beta_{23}x_3] + x_3[\beta_3 + \beta_{33}x_3] \ldots + x_4[\beta_4 + \beta_{44}x_4 + \beta_{24}x_2 + \beta_{34}x_3]$$

At a maximum Hertz stress of 696, 440 psi (assumed to be 700,000), $X_4 = 0$, Table 4. If, therefore, the B_{10} lives are evaluated at a stress of 700,000 psi, the modified equation reduces to:

$$B_{10}\text{ life (700,000 psi)} = \beta_0 + x_1[\beta_1 + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3] \ldots + x_2[\beta_2 + \beta_{22}x_2 + \beta_{23}x_3] \ldots + x_3[\beta_3 + \beta_{33}x_3]$$

which is limited to: (*a*) Maximum Hertz stress of 700,000 psi; (*b*) hardnesses greater than 54 R_c; (*c*) grain size larger than ASTM 10; (*d*) surface finishes better than 10 rms, because of the statistical data used in obtaining the equation.

Values of X_1, X_2, and X_3 are tabulated in Table 6.

Table 6 Conversion table

R_c	X_1	$G.S._1$	X_2	$S.F._2$	X_3	Stress x 10^{-3}	X_4
50	-2.177130	1	-1.892105	1	-1.701967	600	-1.001894
51	-1.907683	2	-1.573354	2	-1.501753	610	-2.69335
52	-1.63823	3	-1.25460	3	-1.298538	620	-2.38177
53	-1.368789	4	- .935852	4	-1.095324	630	-2.070183
54	-1.09934	5	- .617101	5	- .892110k	640	-1.75859
55	- .82989	6	- .298350	6	- .688896	650	-1.44.7009
56	- .560469	7	- .020400	7	- .485682	660	-1.13542
57	- .2910025	8	.339151	8	- .282467	670	- .82383
58	- .0215574	9	- .657902	9	- .079253	680	- .5122k8
59	.248891	10	.97665	10	.1239606	690	- .2006619
60	.517337	11	1.2952o	11	.327174	700	.110924
61	.78678	12	1.61144	12	.5303k1	710	.422511
62	1.05623	13	1.9329	13	.73360	720	.734098
63	1.32567	14	2.2516	14	.93681	730	1.0k5685
64	1.59512	15	2.570k0	15	1.1k003	740	1.357275
65	1.86457	16	2.88915	16	1.3k32k	750	1.6688
66	2.13401	17	3.20791	17	1.5k6k0	760	2.036531
67	2.40346	18	3.526610	18	1.7k967	770	2.29203
68	2.672912	19	3.8k512	19	1.95288	780	2.60362
69	2.942k35	20	4.16k163	20	2.156102	790	2.91520
70	3.211805	21	4.k82914			800	3.2267
71	3.k81252	22	4.80166				
72	3.75069	23	5.120k16				
73	4.020k66	24	5.4391[illegible]				
74	4.28959	25	5.757918				
75	4.55903						
76	4.82848						
78	5.097933						
79	5.63682						
80	6.9062738						

Quantitative Results of Modified-B_{10}-Fatigue-Life Equation

The graphical representation of the effects and interactions of hardness, grain size, and surface finish upon the B_{10} fatigue life on M-50 bearing material are shown in Figs. 5 to 17. Figs. 5 to 9 represent the interaction effects of grain size and surface finish upon fatigue life at different levels of surface finish. Figs. 10 to 14 represent the interaction effects of surface finish and hardness upon life at different grain-size levels. Figs. 15 to 17 represent the interaction effects of surface finish and grain size upon fatigue life at different levels of hardness. These results are graphically summarized in the grain size, surface finish, and hardness "sensitivity curves," Figs. 18 to 21.

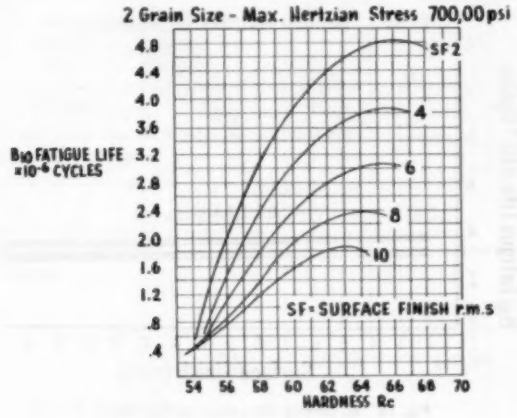


Fig. 5 B_{10} fatigue life versus hardness

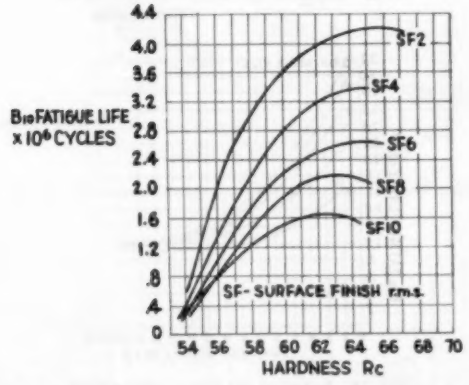


Fig. 6 B_{10} fatigue life versus hardness; grain size 4, maximum Hertzian stress 700,000 psi

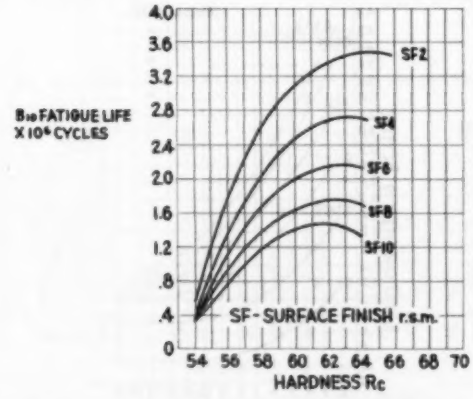


Fig. 7 B_{10} fatigue life versus hardness; grain size 6, maximum Hertzian stress 700,000 psi

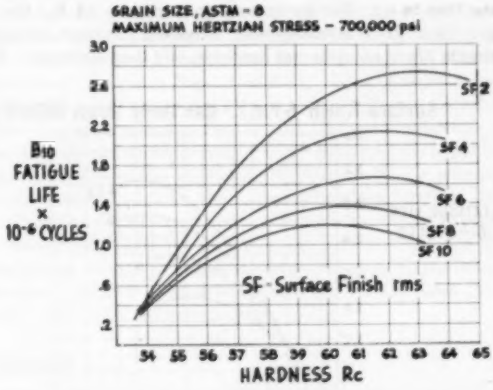


Fig. 8 B_{10} fatigue life versus hardness

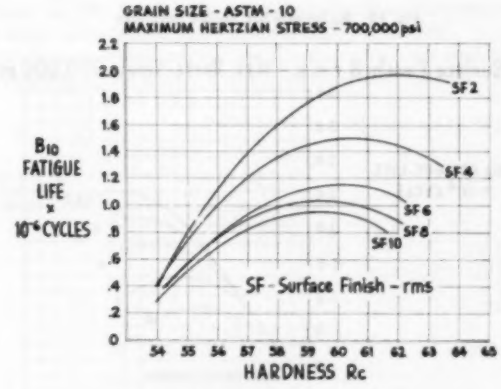
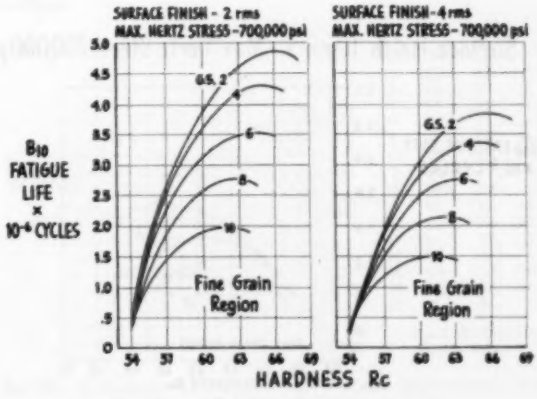


Fig. 9 B_{10} fatigue life versus hardness



Figs. 10 and 11 B_{10} life versus hardness

Discussion of Qualitative Results Obtained From Modified-B_{10}-Fatigue-Life Equation

Although the quantitative results obtained from the formula are based upon a maximum Hertz stress of 700,000 psi, the qualitative results indicate the following effects of hardness, grain size, and surface finish upon the B_{10} life of M-50.

Effect of Hardness Upon the B_{10} Fatigue Life of M-50 (For Hardnesses Greater Than 54 R_c). For hardnesses greater than 54 R_c, the B_{10} fatigue life of M-50 increases with increasing hardness, reaches a maximum life at an optimum hardness, and then decreases. This effect is shown in Figs. 5 to 14. This optimum hardness is not a constant but varies with surface finish and grain size. This effect is shown in Fig. 22.

Effect of Grain Size Upon the B_{10} Fatigue Life of M-50 (For Grain Sizes Larger Than ASTM 10). For grain size larger than ASTM 10, the B_{10} fatigue life increases as the grain size increases. The B_{10} life approaches a constant maximum value at the large grain sizes

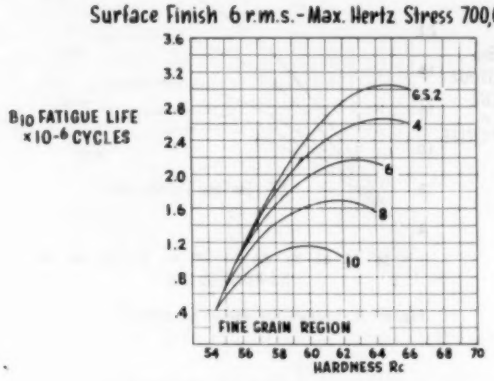


Fig. 12 B_{10} fatigue life versus hardness

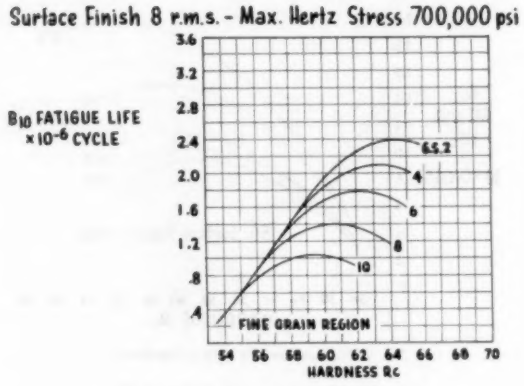


Fig. 13 B_{10} fatigue life versus hardness

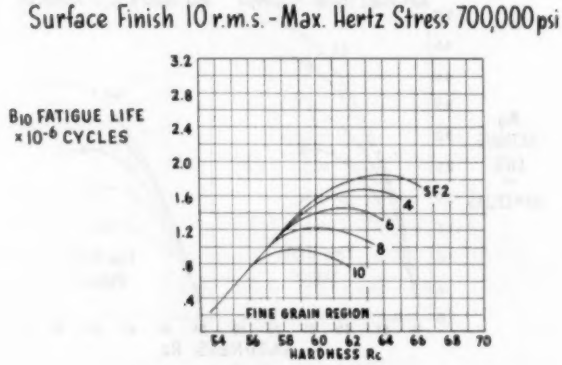


Fig. 14 B_{10} fatigue life versus hardness

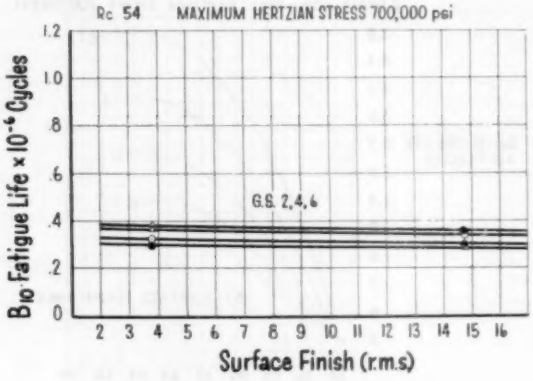


Fig. 15 B_{10} fatigue life versus surface finish

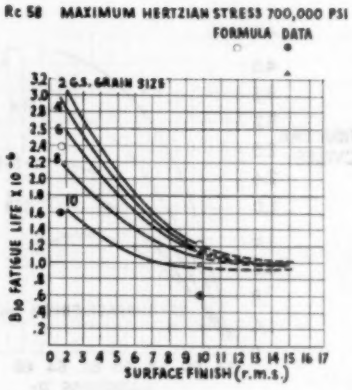

Rc 58 MAXIMUM HERTZIAN STRESS 700,000 PSI
FORMULA DATA
2 G.S. GRAIN SIZE
4
6
8
10
B10 FATIGUE LIFE x 10-6
0 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
SURFACE FINISH (r.m.s.)

Fig. 16 B_{10} fatigue life versus surface finish

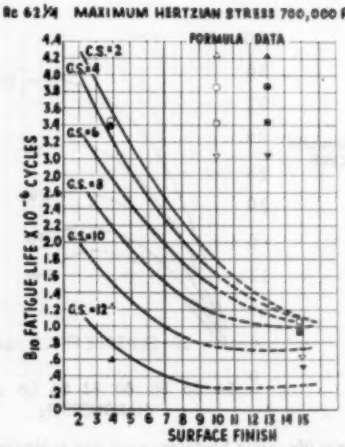


Fig. 17 B_{10} fatigue life versus surface finish

and decreases rapidly as the grains become smaller. This effect is shown in Fig. 20.

Effect of Surface Finish Upon the B_{10} Fatigue Life of M-50 (for Surface Finishes Better Than 10 rms). For surface finishes better than 10 rms, the B_{10} fatigue life increases as the surface finish improves. This life increases steadily with improving surface finish and approaches a minimum at a surface finish of 10 rms. This effect is shown in Fig. 21.

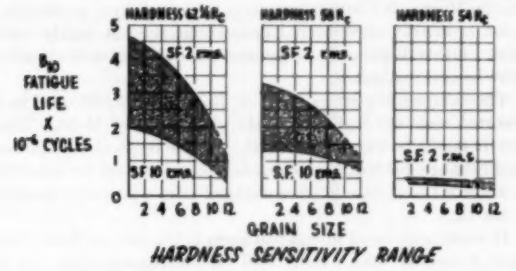


Fig. 18 B_{10} fatigue life versus grain size for various hardness and surface finishes

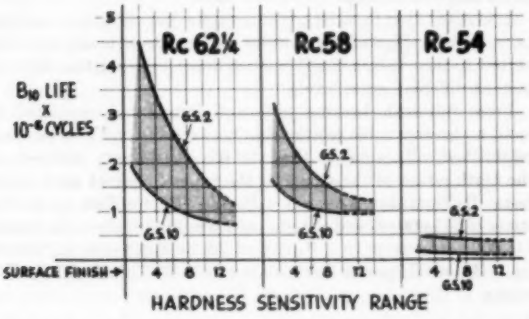


Fig. 19 B_{10} life versus finish for various hardnesses and grain sizes

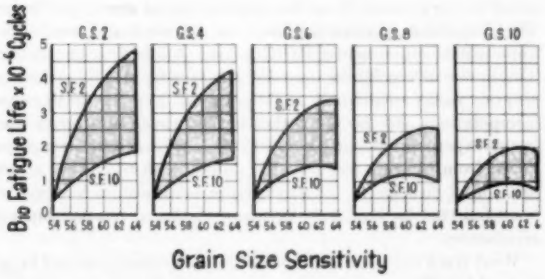


Fig. 20 B_{10} versus hardness for various grain sizes and surface finish

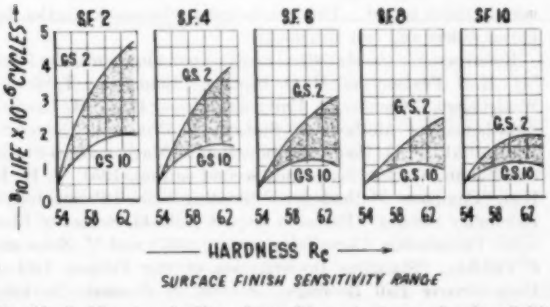


Fig. 21 B_{10} life versus hardness for various grain sizes and surface finish

Interaction Effects of Grain Size, Surface Finish, and Hardness Upon the B_{10} Fatigue Life of M-50 (for Hardness > 54, Surface Finish < 10 rms, Grain Size > ASTM 10). Within the limiting regions of each variable mentioned above, the interaction effects between the variables may be best explained in terms of the "sensitivity" curves. These curves represent a measure of the sensitivity of fatigue life upon two of the variables at different levels of the third variable. The most significant results can be seen from the "hardness-sensitivity" curves in Figs. 20 and 21. These curves show that at the high hardness level of 62¼ R_c, the B_{10} life is extremely sensitive to grain size and surface finish. The area under this curve is a measure of this sensitivity. This sensitivity decreases at the intermediate hardness level of 58 R_c, and the B_{10} life becomes insensitive to variation in grain size and surface finish at the low hardness level of 54 R_c. Results obtained from the grain-size and surface-finish sensitivity curves do not exhibit this same drastic interaction effect, although decreased ranges in predicted life occur as the grain sizes decrease and the surface finish becomes poor. This effect is shown in Figs. 18 and 19.

Summary

The effects of hardness, grain size, and surface finish upon fatigue life of M-50 are evaluated within certain ranges of each variable. The experimental and computed results compare favorably at the various hardness, surface-finish, and grain-size levels, Fig. 23. It should be noted, however, that this comparison is based upon limited data, which is highly statistical. In addition, further approximations are introduced into this analysis by converting all the test data to a stress level of 700,000 psi.

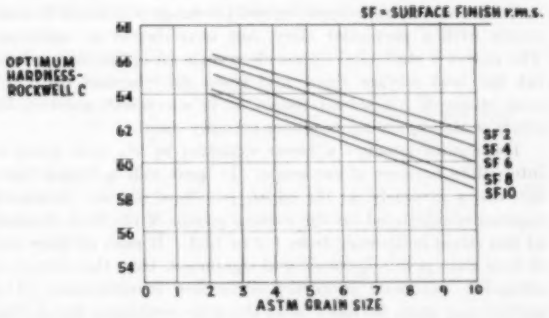
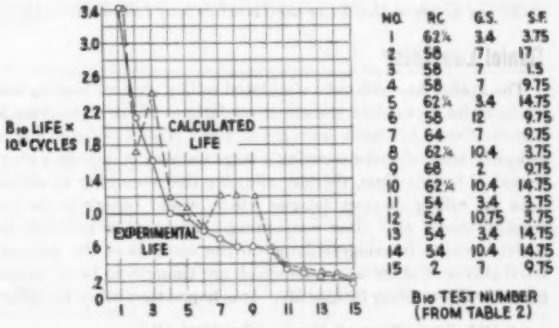


Fig. 22 Effect of grain size and surface finish upon optimum bearing hardness



NO.	RC	G.S.	S.F.
1	62¼	3.4	3.75
2	58	7	17
3	58	7	1.5
4	58	7	9.75
5	62¼	3.4	14.75
6	58	12	9.75
7	64	7	9.75
8	62¼	10.4	3.75
9	68	2	9.75
10	62¼	10.4	14.75
11	54	3.4	3.75
12	54	10.75	3.75
13	54	3.4	14.75
14	54	10.4	14.75
15	2	7	9.75

Fig. 23 Comparison of computed B_{10} lifes and experimental results (maximum Hertzian stress 700,000 psi)

Acknowledgment

The author wishes to acknowledge the contribution of S. N. Miller and E. N. Bamberger of FPL Metallurgical Laboratory and J. A. Zoellner of the General Engineering Laboratory who contributed to both the experimental work and the evaluation of results given in this paper.

DISCUSSION

Thomas L. Carter and William J. Anderson[4]

The author presents some very interesting data on the effects of hardness, surface finish, and prior austenite grain size on rolling fatigue. The effect of hardness, in particular, on rolling fatigue has been the subject of several investigations with both simplified fatigue tests and with full scale bearing tests. In unpublished work completed several years ago, a bearing manufacturer found that the fatigue life of SAE 52100 bearings increased with hardness to the highest hardness tested (Rockwell C 64). In work at the NASA Lewis Research Center, tests in the fatigue spin rig and the Five Ball Tester on M-50, Halmo, M-1, and WB-49 alloys indicate that for each alloy life improves with hardness. In no instance was there any significant dropoff in life with increasing hardness. Prior austenite grain size was on the order of 6 to 8 and surface finish approximately 1 to 2 microinches rms for all four alloys. Maximum hardnesses tested were as follows:

M50	63.5 Rc
Halmo	62
M-1	66
WB-49	68

With all of these data it was noted that the rate of life increase fell off with increasing hardness. Therefore, if a second degree curve were to be fitted to any of these data, it would predict an optimum hardness, perhaps beyond the range of attainable hardnesses with a particular alloy, but nevertheless an optimum. The author's statistical approach reveals an optimum hardness (at the best surface finish and grain size) beyond the hardness range of his actual tests. It is somewhat questionable whether this optimum hardness actually exists.

The author assumes a linear variation in life with stress to interpolate between stress levels. In most rolling fatigue tests, life varies inversely as the ninth power of stress. Stress-life exponents calculated for the various groups of life data obtained at two stress levels vary from 1.2 to 13.3. If each of these sets of four data points is considered significant, then this scatter in stress-life exponent is worthy of further investigation. This scatter may stem partially from the wide confidence bands that result when only 4 data points are used to form a Weibull plot.

Although, as the author states, the procedures used in the statistical analysis are beyond the scope of the paper, a reference should be given so that interested readers may follow the analysis.

Daniel Lundquist[5]

The author has utilized a simulated rolling contact bearing test rig in order to explore the effect on fatigue life of variations in surface finish hardness and grain size. The simulated rig is uniquely adaptable for securing a large number of data in a short period of testing time, thereby allowing the researcher to obtain data on rolling contact fatigue which would otherwise be extremely costly and time consuming. The author and his associates are to be congratulated on the analysis of the inter-related effects of three variables which are thought to be of importance in determining fatigue life. It is hoped that other metallurgical and dimensional variations can be studied at a later date.

The effect of surface finish on rolling contact fatigue life has been the subject of investigation in full scale tests on ball and roller bearings. These tests have been conducted at compressive stress levels considerably below those reported in this paper. At these levels, tests on 35 millimeter bore bearings conducted at Marlin-Rockwell Corporation have revealed that variations in finish ordinarily encountered in bearings are not highly significant. Ground, ground and polished, honed, and etched surfaces have been investigated.

The compressive stress levels of 732,000—645,000 used in the reported tests are higher than the yield point of M-50. Therefore it might be expected that the surface finish of thes pecimen could change due to a burnishing action. It would be interesting to know if the surface finishes changed after a period of operation in the rig.

It would also be of interest to learn if the surface finish values were measured both axially and circumferentially on the test bars and rollers. It is possible to conceive of a surface which would contain circumferential grooves. Such a piece might have a good finish measured circumferentially, but would cause high stress concentrations in the contact area.

It is noted that in order to compare the test results on an equal basis that the B_{10} and B_{50} lives for various stress levels were converted to new values based upon a linear interpolation between tests at different stress levels.

From full scale bearing tests and from other simulated rig tests, an inverse cubic relationship between life and load has been established. This means that the life should vary inversely as the ninth power of the stress for the type of contact used in this test. By linear interpolation, a B_{50} life of 4.4 million cycles at a stress of 645,000 psi and a B_{50} life of 1.5 million cycles at a stress of 745,000 psi convert to a B_{50} life of 2.8 million cycles at 700,000 psi. By ninth power interpolation, the B_{50} life is 3.2 million cycles.

Author's Closure

The relationship between stress and life in bearings has been noted to vary inversely as the ninth power of stress, not linear. This important relationship has been observed and verified by unpublished experimental RC Rig data obtained by the author. The linear interpolation used for the benefit of this statistical analysis yields results only slightly varied from the ninth power interpolations and, for the rather narrow range of stresses used in this experimental work, is within the experimental error introduced from the test results. We do not wish to convey an expression of disbelief in the ninth power relationship, however, to point out that for this analysis, use of the linear relationship was satisfactory.

Wear track surface finish, measured both transverse and longitudinal to rolling, was found to improve slightly during running. For example, the rms 17 surface finish was rms 10-12 after long running. Similar improvements were observed for all levels of surface finish tested. This results from a burnishing action between rollers and test specimens.

References for the statistical analysis techniques may be found in: E. S. Pearson and H. O. Hartley, "Biometrika Tables for Statisticians," Cambridge University Press, 1958; H. Cramer, "Mathematical Methods of Statistics," Princeton University Press, 1951; P. G. Hoel, "Introduction to Mathematical Statistics," John Wiley & Sons, Inc., second edition, 1954; J. H. K. Kao, "Computer Methods for Estimating Weibull Parameters in Reliability Studies," Research Report #EE344, School of Electrical Engineering, Cornell University, 1957; and M. Zelen and J. Fieblein, "Statistics Investigation of the Fatigue Life of Deep Groove Ball Bearings," *Journal of Research*, National Bureau of Standards, vol. 57, 1956, p. 273.

[4] NASA, Lewis Research Center, Cleveland, Ohio.

[5] Assistant Chief Engineer-Administrative, Marlin-Rockwell Corperation, Jamestown, N. Y.

The Importance of Spinning Friction in Thrust-Carrying Ball Bearings

G. S. REICHENBACH
Assistant Professor of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. Assoc. Mem. ASME

Experimental work was done rolling balls on flat plates and in V-grooves at loads and contact angles corresponding to usual thrust-bearing practice. It is shown that the spinning action of the ball with respect to the race should account for the major part of the over-all friction of a thrust-carrying ball bearing. Variables studied included contact angle, conformity, load, lubricant, and temperature. The results have been correlated and shown to follow theoretical predictions.

Introduction

THE coefficient of friction of a ball bearing is very small, of the order 0.001 to 0.003. Yet even this small value is of great concern to designers in many applications. The main thrust bearing of a jet engine which is subject to extreme loads and speeds may absorb as much as 25 hp. The lost power is not so important as the problem of removing this large quantity of energy from the cooling oil. At the other extreme is the problem confronting the designer of gyroscope bearings used in inertial-guidance systems. He has limited power available in his driving motor to overcome the friction and is troubled by navigational errors which are an outgrowth of the bearing friction. While gross values of friction coefficient for many types of bearings are known, the basic details of the motion which lead to the over-all friction force and the relative importance of the various factors are still not completely resolved.

This paper will restrict itself to consideration of ball bearings of the thrust type. Many of the conclusions drawn will be applicable to other rolling-contact situations but will not be discussed here.

If a ball in a high-speed bearing is examined, it is found to be under the influence of several forces, Fig. 1.

In general, because of the centrifugal force on the ball, the points of contact will not be on opposite ends of a ball diameter but instead the outer contact angle α_o will be less than the inner angle α_i. Jones [1][1] has worked out in detail methods for calculating these angles. If the relative angular velocities of a ball are examined with respect to a race, it is found that for a thrust-carrying ball bearing simple rolling motion cannot obtain. As can be seen from the vector diagram of Fig. 2, the ball must have a spinning velocity $\omega_{B/i}$ with respect to the inner race. The condition of outer-race control is illustrated; i.e., pure rolling at the outer race contact, rolling plus spinning at the inner race. Depending upon the relative conformities, speed, and ball size,

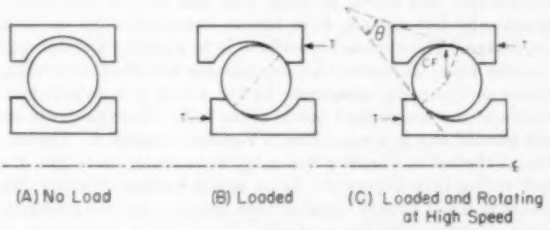


Fig. 1 Forces acting on a ball in a high-speed angular contact bearing

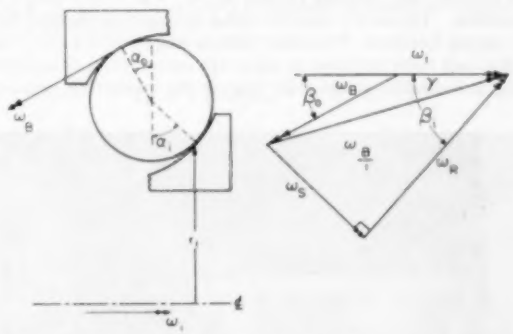


Fig. 2 Angular velocity relationships for a ball in an angular contact bearing

[1] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20-22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 3, 1959. ASME Paper No. 59—Lub-9.

Nomenclature

α_i = inner-race contact angle
α_o = outer-race contact angle
T = thrust load on a bearing
CF = centrifugal force of a single ball
R_o = normal force on a ball at outer-race contact
R_i = normal force on a ball at inner-race contact
ω_i = angular velocity of inner race
ω_B = angular velocity of a ball
$\omega_{B/i}$ = angular velocity of inner race with respect to a ball
ω_S = spinning component of relative velocity
ω_R = rolling component of relative velocity
r = radius of curvature of a race
f = conformity $f = r/D$
μ_S = sliding-friction coefficient in a spinning Hertz contact zone
μ_R = rolling coefficient of friction = F/N
F = force required to roll a ball on a flat plate or in a V-groove
N = normal force at a ball contact
M_R = rolling moment
M_S = spinning moment
D = ball diameter
S = vertical distance between ball contact on upper and lower V-groove
θ = included angle of V-groove (for a flat plate $\theta = 180$ deg)

the motion may be inner-race control, outer-race control, or spinning on both races with neither controlling.

The relative angular velocity between the inner race and ball $\omega_{B/i}$ may be broken into two components, a rolling component and a spinning component. This leads to two sources of energy absorption, moments which resist, respectively, the spinning and the rolling angular velocities. In addition to these two modes of energy absorption, there are several others not directly connected with the rolling action. First in oil churning, a significant contribution when using oil jets to supply the lubricant in high-speed bearings. The use of an air-oil mist may largely eliminate this factor. Also there is friction at the ball-cage contact. The contribution of this has not been measured accurately to date, but the feeling is that the other factors account for most of the observed friction so this must be small.

The work to be reported here concerns itself with only the first two mentioned losses; namely, spinning and rolling friction in the contact zone.

Apparatus

Direct determination of the operating contact angles, ball-spinning velocity, and so forth, is very difficult for real bearings. Shevchenko and Bolan [2] have done this for one case photographically but did not have torque instrumentation on their apparatus. The approach used here is to examine a simple apparatus where the contact angles, spinning velocities, and rolling forces may be easily measured. In Fig. 1 tangent lines have been drawn at the two contact points of the ball. These tangents are not parallel and in a sense form a V-groove of angle θ. The motion of the ball in a bearing is in many ways analogous to that of a ball rolling in a V-groove. In an actual bearing, however, the spinning is not usually equal at both contacts and the conformities may differ at the two contacts.

Experimentally it is fairly simple to roll balls in grooves and measure the resulting friction forces. Fig. 3 shows such an apparatus. The load is applied with a standard mechanical tensile-testing machine. The center plate is reciprocated with a 1-in. stroke, and force required to move the center plate is recorded using a resistance strain-gage ring as the measuring element.

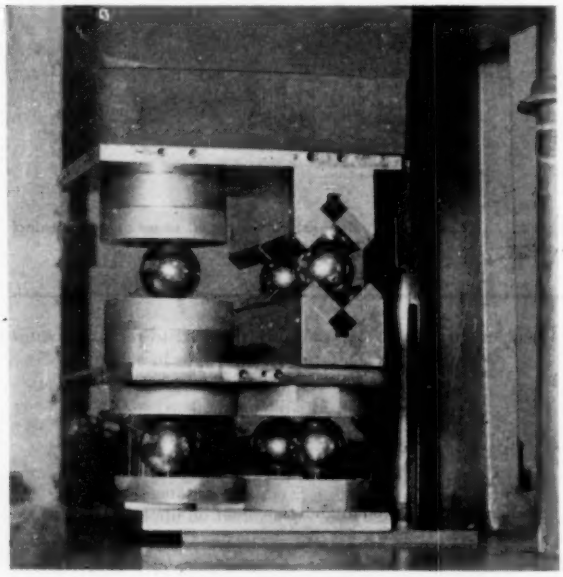


Fig. 3 Test apparatus for V-groove rolling experiments

The force was found to be independent of velocity for rates of 10 cycles per min (cpm) or less. Two tests were run. In the first, the apparatus of Fig. 3 was used without any V-grooves, only flat parallel plates. In the second, a pair of V-grooves was inserted as shown. Only one set of grooves was used to avoid wedging effects caused by possible misalignment of multiple sets of grooves. The apparatus as shown is self-aligning. The balls and rolling surfaces in all cases were 52100 steel, hardness 58-60 Rockwell C. The finish on flat surfaces was lapped, the balls had a standard first-quality finish. The finish on conforming V-grooves was probably the worst of the group, owing to the difficulty of grinding in the range 2–6 μin. rms. Surface finish was found to influence the friction forces slightly, but has not been investigated in detail as yet.

A total of six balls was used, three at each level. The load was assumed to be evenly distributed to each ball. This was accomplished using a 3-in-thick piece of soft rubber between plates at the top of the apparatus to absorb any lack of parallelism between the testing-machine surfaces, or wedging effect due to the motion of the center plate. There was little difficulty in repeating measurements from setup to setup. The testing machine indicated the same total load at any position of the middle plate.

The variables studied were lubricant, temperature, groove angle, conformity at the contact, and stress level.

Experimental Results

Flat-Surface Tests. The first tests used $1^1/_4$-in. balls on flat surfaces and confirmed the results of Drutowski [3]. Defining a rolling-friction coefficient μ_R as in Fig. 4, most values of μ_R were in the range $10^{-4} \rightarrow 10^{-5}$. A rolling friction of this small value may be accounted for entirely by internal hysteresis or damping of the metal. Rolling friction is perhaps more realistically measured in terms of a rolling moment M_R defined as in Fig. 4. A similar total moment may be calculated for balls in V-grooves if the ball diameter is replaced by the distance S as shown in Fig. 7.

Fig. 5 shows a typical curve for flat-plate tests on a log-log plot. Two regions are noted: An apparently elastic region below 150-lb load and a plastic region above 150-lb load. The dividing line between the two regions occurs at a stress of about 190,000 psi for this steel. These results are not reported or analyzed in detail as they do not differ from Drutowski's work. These readings were used only to correct later readings using the V-grooves. It might be noted that in the usual range of viscosity, widely different types of oil or no oil at all gave the same force readings when used on flat plates.

V-Groove Tests. The next group of tests replaced two of the flat surfaces with V-grooves of various conformities. Common practice defines conformity f as:

$$f = \frac{r_R}{D} \tag{1}$$

where r_R is the radius of the race and D is the ball diameter.

Using this definition, a flat plate has ∞ conformity while a typical ball bearing has $52 \rightarrow 54$ per cent conformity.

Fig. 6 shows data for various configurations. All the data have been reduced to a common basis of one ball with two contact points. The rolling moment on a flat plate is almost negligible when compared to balls rolling in various types of V-grooves. This difference is attributed to the spinning action that the ball has as it rolls in the groove.

A free-body diagram of a ball rolling in a V-groove may be drawn as in Fig. 7. The ball is resisted in its motion by both spinning moments M_S and rolling moments M_R. The rolling moments have been determined already from the flat-plate experiments. Moment equilibrium of the ball requires

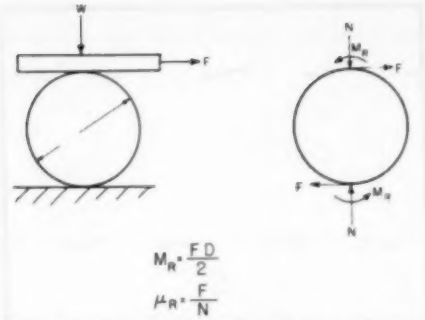


Fig. 4 Manner of defining rolling friction coefficient and rolling moment

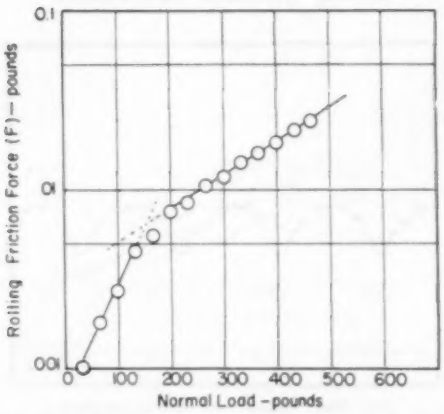


Fig. 5 Rolling force versus normal load for a ball rolling between flat plates—$1^1/_4$-in. diameter ball

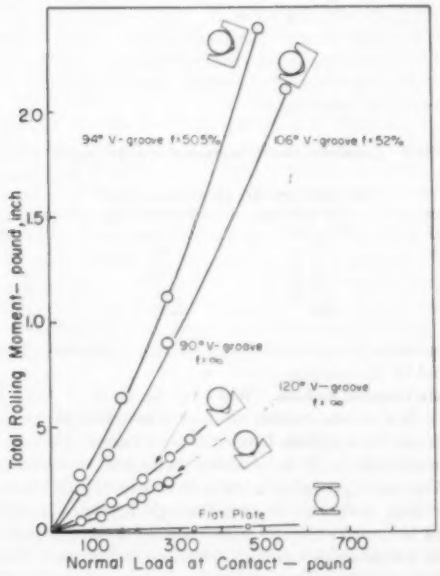


Fig. 6 Rolling moment versus normal load for several configurations—$1^1/_4$-in. balls, MIL 7808 Oil

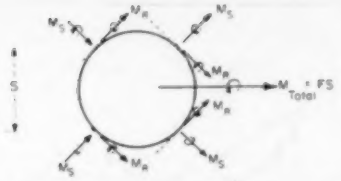


Fig. 7 Free body diagram of moments acting on a ball rolling in a V-groove

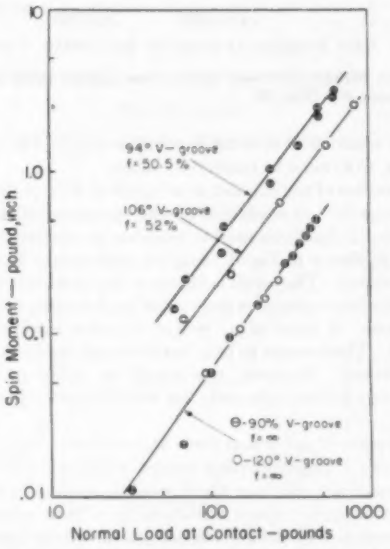


Fig. 8 Spinning moment versus normal load for several configurations—$1^1/_4$-in. balls, MIL 7808 Oil

$$FS = 4\left(M_S \cos\frac{\theta}{2} + M_R \sin\frac{\theta}{2}\right) \tag{2}$$

Using this relation, the spinning moment has been determined for a variety of situations and the results are plotted in Fig. 8. On this log-log plot, all the slopes are approximately $^4/_3$, agreeing quite well with the analysis of Poritsky, Hewlett, and Coleman [4]. They integrated the friction force over the contact ellipse to obtain a spinning moment. A constant spin-friction coefficient μ_S and a normal stress distribution equal to the Hertz distribution were assumed. The result is

$$M_S = \frac{3}{8}\mu_S N a E(k)$$

where $E(k)$ is the complete elliptic integral of the second kind,

$$k = [1 - (b^2/a^2)]^{1/2} \tag{3}$$

a is the major semiaxis of the contact zone, and b is the minor semiaxis of the contact zone. For a circular contact zone ($b/a = 1$), this reduces to:

$$M_S = \frac{3\pi}{16}\mu_S N a \tag{4}$$

If the contact zone is elliptical with $a/b > 7$, less than a 2 per

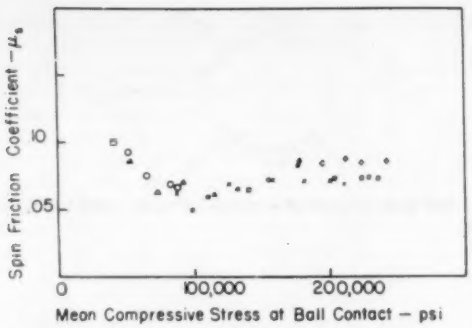


Fig. 9 Spin friction coefficient versus mean contact stress for various configurations, Mil 7808 Oil

cent error results if $E(k)$ is taken equal to unity. For intermediate values, $E(k)$ must be taken from tables.

As the radius of the contact zone varies as $N^{1/3}$, it follows that M_S varies as $N^{4/3}$ as confirmed by the experimental results.

Equation (3) has been used to evaluate the friction coefficient for the data shown in Fig. 6, using the appropriate normal loads and conformity. The result is shown in Fig. 9 where friction coefficient has been plotted versus mean compressive stress in the contact zone. A value of $\mu_S = 0.07$ describes most of the data quite well. There seems to be a trend toward higher values of μ_S at low stresses. However, this should be taken as tentative, since the data become relatively less accurate as the loads are decreased.

Examination of an actual force trace for the V-groove, Fig. 10(*a*), shows a predominating square wave, as expected, plus higher frequency variations which seem to repeat almost exactly from cycle to cycle. These variations have been attributed to local differences in surface finish and material structure. When rolling on plane surfaces, Fig. 10(*b*), these local variations almost obliterate the basic square wave. Local peaks may be ten times the average force. Drutowski [3] has examined these local variations in detail.

The scatter in the friction-coefficient values does not seem to be entirely random. Several series of tests were run using flat-sided V-grooves of various angles. If the friction coefficient is plotted as a function of the spin to rolling angular velocity ratio, a definite trend is noted in all cases. Fig. 11 shows a typical set of tests. Higher coefficients are associated with large relative amounts of spinning. This result seems logical if it is assumed that the friction coefficient depends strongly upon an adsorbed surface film which can be progressively wiped away by the sliding action. If there is a large rolling component, new surfaces are continually brought into contact, while the surfaces which have been sliding are exposed again to the lubricant where the adsorbed film can be renewed. If the rolling component is relatively small, surfaces within the contact area may slide a much greater distance as a result of the spinning before they are again exposed to the lubricant supply.

Lubricant Tests. The use of V-grooves proved to be quite a sensitive method for determining the friction properties of various oils. Fig. 12 shows the force versus normal load for a variety of oils. These values are all reproducible except for the "dry surface." The exact manner and care used in cleaning the surface gave rise to large variations in this result. The data suggest that the running torque of angular contact thrust ball bearings should be strongly dependent upon the type of lubricant.

A calculation was made to see if spinning friction could account for the heat generation of a 240-mm aircraft turbine-type thrust bearing operating at 9000 rpm. The contact angles were evaluated according to the analysis of Jones [1]. The spinning and

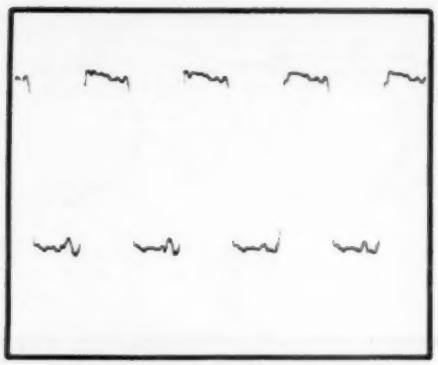


Fig. 10(a) Typical force trace for V-grooves

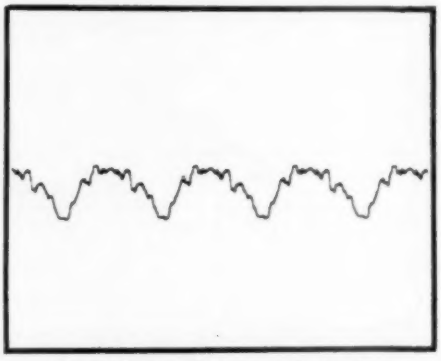


Fig. 10(b) Typical force trace for flat plate

rolling angular velocities were calculated assuming outer-race control. The energy generated is then the product of the angular velocity and its resisting moment. The contribution of all ball contacts is added to give the total energy. Table 1 gives the calculated results compared with experimental values of heat rejection. The experimental values have been adjusted to eliminate the effects of oil churning.[2]

Table 1 Calculated results compared to experimental values

Thrust load, lb	Heat generated by rolling, Btu/min	Heat generated by spinning, Btu/min	Experimental heat generation, Btu/min
2000	7.5	82	90
4000	8.3	159	136
6000	15.6	225	181
10000	36	350	272

The spinning action by itself can account for almost all the heat generated by the bearing.

Variable Temperature Tests. Work by Accinelli [5] and Kingsbury [6] has shown sudden rises or transitions in friction coefficients as the ambient temperature is raised. To see if the same phenomena could be observed here, heaters were attached to the V-groove apparatus in order to control the temperature of the spinning surfaces. A once-through lubricant supply was arranged to avoid problems with thermal decomposition of the oil. The temperature was then cycled up and down with a cycle time of about 2 hr. The maximum temperature was limited by annealing of the 52100 steel surfaces. Fig. 13(*a*) shows the results

[2] These experimental values have been measured by Pratt & Whitney Aircraft using an extrapolation technique to eliminate the effect of churning.

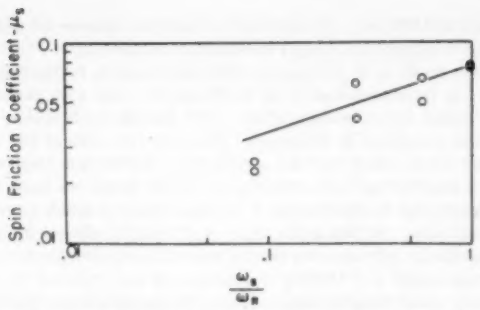


Fig. 11 Spin friction coefficient versus ratio of spinning to rolling, Mil 7808 Oil—$1^1/_4$-in. ball, N = 200 lb

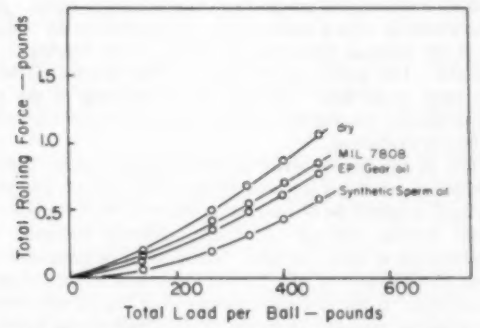
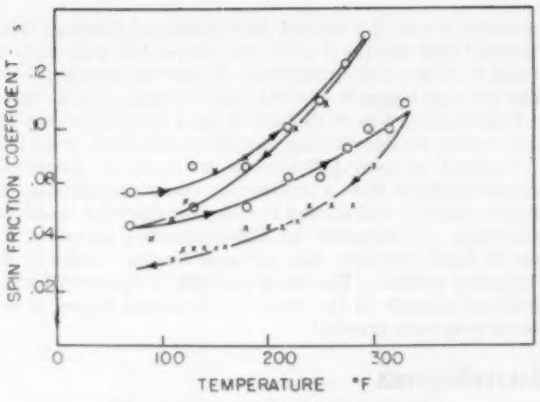


Fig. 12 Rolling force versus normal load for several lubricants θ = 120 deg $1^1/_4$-in. ball

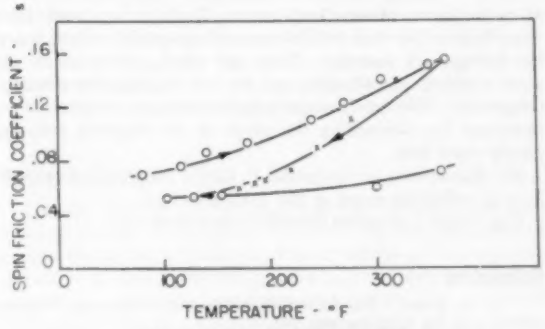
of two such cycles. There is a rise in friction with temperature but the rise becomes less for repeated cycles. However, finally an equilibrium curve is attained as in Fig. 13(*b*). This improvement with a number of cycles is attributed to a gradual buildup of a varnish film with low friction properties. However, the sharp rise in friction still occurs in about the same temperature range. This rise has been attributed to desorption of the lubricant film. If this rise should occur in a high-speed bearing, the whole process may snowball, the friction rise generating more energy resulting in a further temperature rise, and so on, until finally the bearing is destroyed. The critical-temperature-range rise varies greatly for different lubricants. Fig. 13(*c*) shows another example. Fig. 13(*c*) is especially interesting as there is a sharp break in the rising curve at 350 F. At the same time the oil began to smoke and decompose rapidly forming a varnish film on the surface. The reduction in friction is attributed to the boundary lubricating effects of this film. It is doubtful that a film of this nature would form rapidly enough to be helpful in a high-speed bearing.

Conclusions

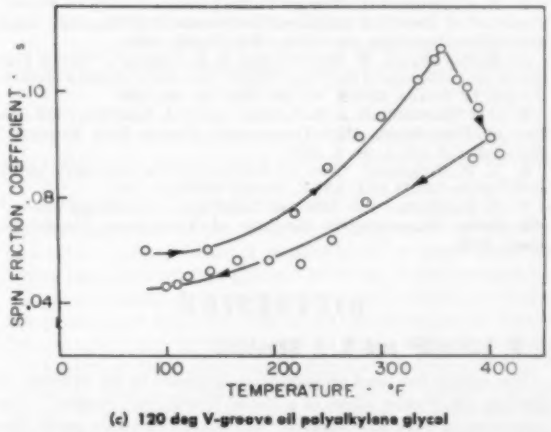
The important conclusions to be drawn here is that in many types of rolling contact the major part of rolling resistance is due to the spinning action of the balls. The so-called rolling-friction moment which has been attributed to hysteresis within the metal is at least an order of magnitude smaller than the spinning moment. While the rolling moment is known to be relatively insensitive to lubrication, the spinning moment being due to actual sliding within the contact zone exhibits all the characteristics of sliding friction; i.e., various oils may cause large changes in the friction coefficient; high temperatures may cause desorption of the lubricating oil film, resulting in abrupt changes in friction moment; the friction coefficient is independent of the contact



(*a*) 120 deg V-groove oil polyalkylene glycol



(*b*) 120 deg V-groove oil Mil 7808



(*c*) 120 deg V-groove oil polyalkylene glycol

Fig. 13 Spin friction coefficient versus temperature, $1^1/_4$-in. ball, N = 260 lb

stress although the moment itself varies as the $^4/_3$-power of normal load due to geometrical considerations. Changes in geometry of the bearing which affect the amount of spinning will result in corresponding variations in the over-all running torque.

These experiments were carried out under low-speed conditions and one might argue that at higher speeds hydrodynamic forces in the contact zone may come into play and complicate the picture. While this is certainly true to some extent, the work of

Accinelli, et al. [5], showed that high-speed bearings, when operated with varieties of lubricants, showed this same sharp increase in friction with temperature. In fact, the silicone oil which had the most favorable viscosity characteristics for maintaining a hydrodynamic film in the face of rising temperature was the first to show the sharp friction increase resulting from desorption.

Inasmuch as many ball bearings are forced by geometrical considerations to have a considerable amount of spinning, designers would be well advised to remember this when specifying lubricants. In particular for high-temperature operation, the use of liquid lubricants does not seem feasible because of the desorption problem. The use of solid-film lubricants either applied mechanically to the surface or chemically formed *in situ* would seem more practical.

Acknowledgments

The author is deeply indebted to the Pratt & Whitney Division of the United Aircraft Corporation for its generous support of bearing and seal research over the past several years. In particular to Messrs. Paul Brown, Richard Shevchenko, and Gino Andreini for their valuable counsel and wealth of indispensable background material. They are much interested in this same problem of ball spin and are investigating its effect on fatigue life. The experimental values of heat generation and the technique for eliminating the effect of oil churning are completely their data.

My thanks also to Kishorelal P. Ghiya for his meticulous efforts in collecting much of the data presented.

Figs. 1 and 2 are taken directly from reference [5].

References

1 A. B. Jones, "The Life of High-Speed Ball Bearings," TRANS. ASME, vol. 74, 1952, pp. 695–703.

2 R. P. Shevchenko and P. Bolan, "A Visual Study of Ball Motion in a High-Speed Thrust Bearing," presented at the SAE Annual Meeting, January 14–17, 1957.

3 R. C. Drutowski, "Energy Losses of Balls Rolling on Plates," presented at Technical Session of Instrument Ball Bearing Conference—New Departure, Cleveland, Ohio, March, 1958.

4 H. Poritsky, C. W. Hewlett, and R. E. Coleman, "Sliding Friction of Ball Bearings of the Pivot Type," *Journal of Applied Mechanics*, vol. 14, TRANS. ASME, vol. 69, 1947, pp. 261–268.

5 G. S. Reichenbach, J. B. Accinelli, and S. J. Beaubien, "Lubrication of High-Speed, High-Temperature Thrust Ball Bearings," *Trans. ASLE*, vol. 1, no. 2, 1958.

6 E. P. Kingsbury, "Heat of Adsorption of a Boundary Lubricant," paper 59AM,3B4, ASLE, Annual Meeting, 1959.

7 K. P. Ghiya, "The Effect of Lubricants on Rolling Friction," MS thesis, Massachusetts Institute of Technology, Cambridge, Mass., 1958.

DISCUSSION

J. B. Accinelli[3] and S. J. Beaubien[4]

The author has used an ingenious approach to the problem of defining the friction effect in a thrust loaded ball bearing. The experimental work was especially well conceived, to the point, and appeared to allow for good control of the significant variables.

At Shell Development we have been aware of the possible effects of ball spin on the performance of very high-speed thrust loaded ball bearings and, therefore, have used such a bearing in our studies of ball bearing lubrication at very high temperature. In our early work, a 20-mm bearing was operated at 40,000 rpm and 200 lb axial load which gave a calculated relative ball spin velocity of 19,000 rpm and a maximum Hertz stress in the contact zone of about 240,000 psi. Under these conditions silicone oil did not support lubrication even at low ambient temperature. One conclusion made at that time was that ball bearing lubrication appears to be predominantly of the boundary type with perhaps a very small hydrodynamic effect. Dr. Reichenbach makes note of this conclusion in his paper. However, in some of our most recent work, which includes speeds up to 60,000 rpm (representing a relative ball spin velocity of 53,000 rpm), we have some evidence that hydrodynamic lubrication plays a much more important role. At this much higher spin velocity silicone fluid will satisfactorily lubricate the bearing provided an adequate flow rate is maintained and bearing clearances are not reduced to zero. In fact, under these conditions it has been possible to operate satisfactorily with a silicone up to 700 F. In the same series of experiments it was found that total bearing friction was the same for silicone, MIL-L-7808C, and a mineral oil if these fluids are compared at temperatures where their viscosities are equal. Since the lubricant in all our experiments was injected at the temperature of the bearing, there was no cooling of the bearing by the lubricant. The failure of the bearing with silicone oil in our early work could have been caused by churning of the more viscous silicone, leading to rapid internal heating of the bearing and loss of clearances.

While at the speeds given in Reichenbach's paper the spin velocities are relatively low, and hydrodynamic effects are probably small, it would be of interest to study in the V-groove apparatus several lubricants of widely different boundary film properties but of *equal viscosity*. These could be compared with respect to coefficient of spin friction at various Hertz stresses or normal load. If the lubricants compared in Fig. 12 are assumed to increase in viscosity from the MIL-L-7808 to the synthetic sperm oil in the order shown, then the effect observed could be explained by an increasing fraction of the load being supported hydrodynamically.

The work at various temperature was most interesting. In our work with *high-speed* ball bearings, friction *decreased* with *increasing* temperature up to a critical temperature for the lubricant (generally above 600 F). We attribute this to a decrease in viscosity which gives lower viscous resistance and churning. Above the critical temperature we found that friction increased steadily. We attribute this to an increase in the portion of the load supported by a boundary film, rather than to desorption.

The author has contributed a very fine piece of basic experimental work in an imperfectly understood area of ball bearing research. The ball spin effect discussed in the paper is not only of interest with respect to lubrication of ball bearings but also should be valuable to those studying ball bearing design, particularly, where bearing fatigue is a critical factor.

G. J. Andreini[5]

The over-all complexity of ball thrust bearings has made prediction of performance in new applications a matter of "engineering guesstimating." In all cases, these predictions are based on empirical data obtained by experience with past applications. Extrapolation of these data to any great extent has led to many unsatisfactory applications. For these reasons, there is a growing need for a better understanding of ball thrust bearing operation.

One method of tackling the problem is to break it down so that solutions are obtained for pertinent parts of the problem. Applying this philosophy, Pratt & Whitney Aircraft sponsored this work under Dr. Reichenbach as one part of the over-all understanding of ball thrust bearing operation. This work, in combination with the work of Mr. A. B. Jones on dynamics in

[3] Shell Development Company, Emeryville, Calif.

[4] Supervisor, Research, Shell Development Company, Emeryville, Calif. Mem. ASME.

[5] Project Engineer, Pratt & Whitney Aircraft Division, United Aircraft Corporation, West Hartford, Conn.

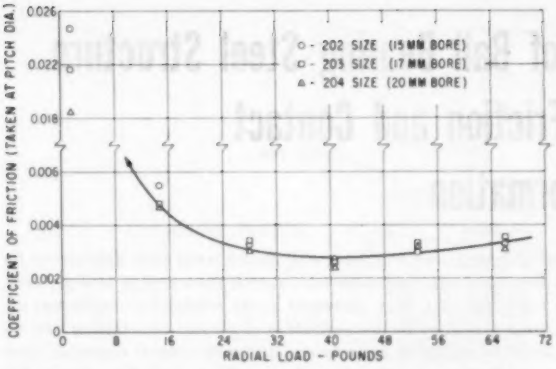


Fig. 14 Breakaway friction coefficient as a function of radial load

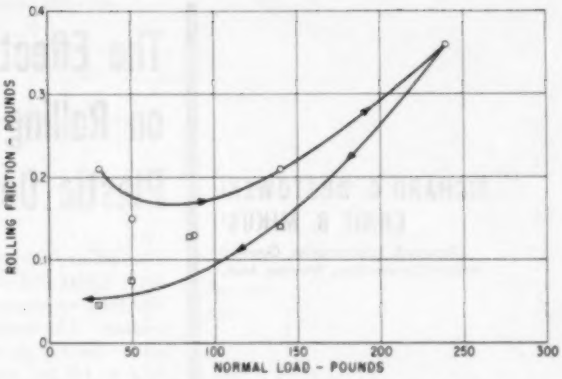


Fig. 15 Rolling friction force as a function of load

ball thrust bearings, represents a major step toward understanding and predicting heat generation in bearings.

In addition to the pronounced increase in friction with increasing spin in the contact area, work now in progress at Pratt & Whitney Aircraft has shown a definite relationship between spin and fatigue life. This testing has shown a decided decrease in life with an increase in spin in the contact area and suggests that bearing geometry may be optimized to reduce spin while maintaining reasonable surface stresses due to applied load.

It is hoped that other investigators will be encouraged by the results reported here to work on the many other significant parts of the over-all problem, such as effects of lubricants and their properties, effect and formation of hydrodynamic films, and dynamic instabilities as related to ball bearings, so that the combination of all of these solutions will eventually lead to a complete understanding and prediction of ball thrust bearing performance.

Paul Lewis[6]

The author is to be congratulated for work in the field of ball bearing friction which has not received sufficient attention. These comments are largely intended to offer supplementary information or to elicit the author's comments on other possibilities.

The first comment is on the relation between rolling friction force as a function of normal load as shown in Fig. 5. This shows a lower friction increase with load in the plastic region (above 150 lb) than in the elastic region. Because this was obtained with flat specimens and balls, the friction should be mainly due to hysteresis. It would be expected that the steeper slope should be in the plastic region.

Fig. 9 shows the sliding friction coefficient as a function of mean compressive stress. Higher values of friction coefficient are suggested at lower stress levels. Fig. 14 shows the breakaway friction torque (based on the bearing pitch diameter) of a deep groove ball bearing as a function of radial load. Here also the over-all coefficient is very high at light loads.

Fig. 13(*a*) shows the variation of the coefficient of friction with temperature and cycling. The reduction in friction is attributed to the formation of varnish. Some of our own work on a similar configuration has yielded Fig. 15, showing the variation of friction with load. The curve is similar to 13(*a*) and was attributed to initial plastic flow and establishment of an equilibrium track. This implies that 13(*a*) might be caused in part by a softening with temperature and plastic flow of the track.

The final comments refer to the temperature dependence of friction and lubricant. Recent work with a once through system has indicated certain temperature sensitive regions. The region with the silicones occurred at relatively low temperature, in the range of 250 F. Further testing at 600 F with the same fluid gave excellent performance. Very small quantities of fluid were used, and it was found that quantity of flow was very critical. Other work at 700 F showed a definite relationship of bearing wear with the flow rate.

[6] Engineer, Ball Bearings, General Engineering Laboratory, General Electric Company, Schenectady, N. Y. Assoc. Mem. ASME.

Author's Closure

The comments of Accinelli and Beaubien regarding high-speed bearings are very interesting. The work reported here certainly cannot explain all their results. Viscous effects were unimportant in this work while for 60,000 rpm bearings viscosity must be very important. However, room-temperature viscosity cannot be taken as an accurate indication of behavior in bearings. It is well known that pressure, temperature, and shear rate can alter viscosity by many orders of magnitude. However, until actual numbers are known for these variations, one can only speculate as to the reasons for the high-speed behavior noted.

Oils of equal viscosity under ambient conditions but of different boundary properties were tested. Addition of 1 per cent stearic or palmitic acid to a white mineral oil reduced the friction forces by 50 per cent. Another indication of the absence of viscous effects in these tests was the absence of any speed effects. The same friction forces were noted at 6 strokes per hr and 10 strokes per min.

The apparently backward behavior of rolling friction in the plastic zone noted by Mr. Lewis is due only to a poor choice of co-ordinate paper. Use of semilogarithmic paper gives the wrong impression at first glance. However, if the numbers are examined carefully, it is seen that increasing the normal load in the plastic zone from, say, 400 to 500 lb, increases the friction force by about 0.01 lb, while a similar increase in the elastic zone, say from 50 to 150 lb, increases the friction force only 0.005 lb.

Fig. 14 by Mr. Lewis shows similar friction behavior at low loads. However, the explanation here must be different, since there should be no spinning in a radially loaded bearing. Perhaps the rise in coefficient can be attributed to cage friction. Even at zero applied load, the cage will cause a significant friction moment.

The gradual drop of friction force with repeated cycles, as in Fig. 13 and Mr. Lewis' Fig. 15, may indeed be due to plastic flow rather than varnish formation. Actually, in some tests indentations several wave lengths of light deep were found in the surface after running.

The results Mr. Lewis' reports with silicone oil sound very similar to the results mentioned by Accinelli and Beaubien. I can give no really good hypothesis for the complicated dependence on temperature, flow rate, and oil type.

The Effect of Ball Bearing Steel Structure on Rolling Friction and Contact Plastic Deformation

RICHARD C. DRUTOWSKI
ERNIE B. MIKUS[1]
Research Laboratories, General Motors Corporation, Warren, Mich.

The rolling friction, contact plastic deformation, and elastic limit were determined for SAE 52100 steel structures with retained austenite contents from zero to 18.4 per cent. The force necessary to roll a ball on a plate decreased as the retained austenite was decreased. The contact stress necessary for the initiation of plastic deformation and the elastic limit of the material increased as the retained austenite content decreased from 18.4 to 3.9 per cent. No further change occurred when the retained austenite was reduced to zero. The extent of plastic deformation at very high contact stresses was reduced by the presence of retained austenite contents up to at least 7.4 per cent. These observations were applied to the problem of selecting the best steel structure for an instrument ball bearing.

Introduction

WHILE IT IS GENERALLY ACCEPTED that ball bearings have very low friction torque, even this low torque is of importance when the ball bearings are used in instruments of high accuracy. The friction torque of a ball bearing is the result of a combination of sliding and rolling friction components. This paper does not consider any of the sliding friction effects that may arise either in the contact area between the ball and race, or between the ball and separator, but examines only the effect of the metallurgical structure of SAE 52100 steel on pure rolling.

For simplicity, the experimental arrangement chosen is that of a ball rolling between two flat plates. When a ball rolls on a flat plate the energy loss is that associated with hysteresis losses of the elastically strained metals [1, 2].[2] The contribution of differential slip in the contact area is negligible as is indicated by the fact that lubricants have little effect on the magnitude of the rolling resistance. For balls rolling in deep grooves the differential slip becomes more important, but this effect is not examined in this paper. The determination of the resistance to rolling for the experimental arrangement employed is a measure of the effect of material hysteresis on instrument ball-bearing performance.

The structures of SAE 52100 steel that were selected for this investigation have retained austenite contents from zero to 18.4 per cent. The interest in these structures originated in the fact that for this range the hardness reaches a maximum of Rc 64 at about 4 per cent retained austenite and drops to a hardness of Rc 58 at zero and 18.4 per cent retained austenite. While the stability associated with the zero retained austenite structure is desirable [3], the effect of the reduction of hardness on rolling friction characteristics and on the resistance of the material to plastic deformation remained to be determined.

Procedure

Test Samples. The plates and balls used for this investigation were made from high quality SAE 52100 ball bearing steel. Four heat-treatments were used to produce significant variations in structure and physical properties:

A Austenitize $^1/_2$ hr at 1550 F, oil quench 90–100 F, temper 1 hr at 500 F.

B Austenitize $^1/_2$ hr at 1550 F, oil quench 90–100 F, subcool $^1/_2$ hr at −321 F + $^1/_2$ hr at 250 F, cycled 5 times.

C Austenitize $^1/_2$ hr at 1550 F, oil quench 90–100 F, temper 10 hr at 250 F.

D Austenitize $^1/_2$ hr at 1800 F, oil quench 90–100 F, temper 1 hr at 350 F.

After the heat-treatment the surfaces were ground and polished to a surface roughness of less than one microinch in preparation for the rolling friction runs. Retained austenite determinations were made on these surfaces using a modified x-ray diffraction technique based on the work of Averbach and Cohen [4].

Elastic limit in tension was determined on specimens having a 0.252-in-diam reduced section using a residual strain technique described by Muir, et al. [5].

For determinations of the rolling characteristics and of the contact stress necessary for the initiation of plastic deformation, the steel balls of a given structure were rolled between plates of the same structure. Additional data on the contact stress necessary for the initiation of plastic deformation were obtained by rolling a $^1/_4$-in. diam cemented-carbide ball, with an elastic modulus of 72×10^6 psi, on each of the plate structures, and data on the extent of plastic deformation at high stresses were obtained by rolling $^1/_2$ and $^1/_4$-in-diam balls of cemented-carbide on each of the plate structures.

Rolling Friction Apparatus. The rolling friction apparatus is shown in Fig. 1. The lower sample plate is located in a holder which is supported by a high pressure oil film and is elastically restrained by a pair of rolling force beams (only one of which is shown in the figure). The upper sample plate is located in a holder which is attached to the free end of the normal force beams. Moving the carriage along the straight ways by a hydraulic ram (not shown) causes the ball to roll between the plates. The resistance to rolling is measured by the deflection of the rolling force beams, and the load is maintained and measured by the deflection of the normal force beams. The apparatus and the technique for measuring rolling force are described in detail in a previous publication [2].

[1] At present, Convair, Astronautics, San Diego, Calif.

[2] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, June 10, 1959. Paper No. 59—Lub-2.

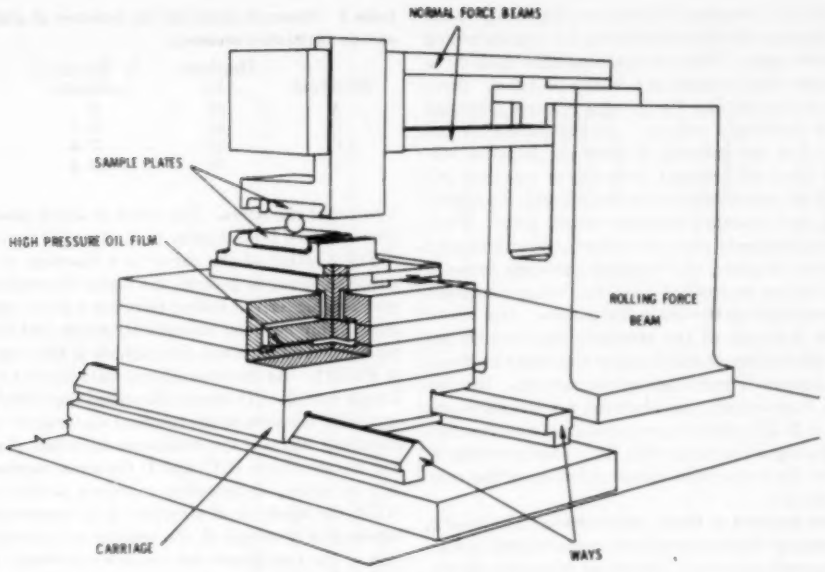


Fig. 1 Rolling friction test apparatus

A rolling friction run consisted of rolling the ball back and forth between the two plates and over the same track until equilibrium was established and then determining the average rolling force for five pairs of passes. Equilibrium conditions were attained within five pairs of passes for the samples used. For all the runs the velocity of the ball center with respect to one of the plates was held constant at 0.03 in. per second and the length of the track was $1^1/_2$ in. The values of the rolling force were calculated for the contact of the ball with only one of the plates. Average rolling force was determined as a function of contact stress for each structure.

Plastic Deformation. In addition to determining the rolling friction characteristics of the four structures, observations were made of the initiation and extent of plastic deformation in the track. At small contact stresses no plastic deformation was evident but as successively higher contact stresses were applied, the plates developed a groove that was identified by visual observation. As the contact stress increased the extent of plastic deformation grew larger. The resulting grooves were measured with a surface profile instrument and the extent of plastic deformation was determined for various contact stresses on the four structures.

Table 1 Some physical properties of various hardened and tempered 52100 steel structures

Structure	Elastic limit, psi	Hardness Rc	% Retained austenite	Dimensional stability microin/in. change in 1000 hr at 165 F and 20,000 psi
A	123,000	58	0	7–18
B	127,000	64	3.9	65–89
C	58,000	62	7.4	95–110
D	...	59	18.4	...

Results and Discussion

Physical Properties. The physical properties that were determined for these four structures of 52100 steel are presented in

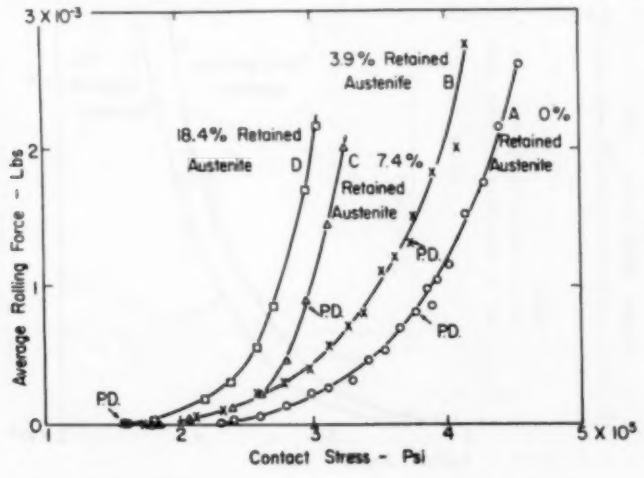


Fig. 2 Average rolling force versus contact stress for various structures of SAE 52100 steel. Ball, 1/2-in. diam, of same structure as plates.

Table 1. The influence of retained austenite on the ability of the material to resist microplastic deformation can be seen by noting its effect on the elastic limit. With retained austenite levels up to 3.9 per cent the elastic limit remains at a maximum value. However, when the structure contains 7.4 per cent retained austenite the elastic limit is drastically reduced. At the same time the dimensional stability of the material is affected. Superior stability is achieved when all retained austenite is removed [3]. With as little as 3.9 per cent retained austenite present, the dimensional stability of the structure becomes rather poor. While these characteristics change in relation to the retained austenite in a rather consistent manner, the hardness variation between structures changes in an anomalous way, not being consistent with either the elastic limit or the retained austenite. One would expect that as the hardness of the structure increases its resistance to plastic deformation would increase also, since hardness is a measure of a material's resistance to penetration. But for 52100 steel a high elastic limit was observed for the structure having a hardness of 58 Rc, while a low elastic limit was observed for the structure having a hardness of 62 Rc. This inversion of hardness and elastic limit has been observed before in the case of plain carbon steels [5].

Rolling Friction and Initiation of Plastic Deformation. The rolling friction characteristics of the four structures are compared in Fig. 2, where the average rolling force is plotted as a function of the contact stress. The contact stress is the maximum compressive stress in the contact area and is given by

$$q_0 = \frac{3}{2}\frac{P}{\pi a^2},$$

in which P is the load and a is the radius of the circular contact area as determined from the Hertzian equations [6]. In each case the data were obtained by using a $^1/_2$-in-diam ball of the same heat-treatment as the plates. Each curve can be divided into three portions. The first is characterized by values of the rolling force so small that they cannot be measured with any accuracy with the apparatus; this portion extends from zero stress to a certain threshold stress that is dependent on material properties. The second portion of the curve starts at this threshold stress and the rolling force increases linearly with the volume of stressed material until a point is reached at which plastic deformation starts; the quantitative relationship between the rolling force and the stressed volume is discussed in another paper [7]. The third portion of the curve covers the region in which the contact stresses are sufficiently high to cause plastic deformation during initial rolling; in this region the rolling force increases as a power of the stress. The stress at which plastic deformation is first observed is marked by the letters PD for each of the curves.

Table 2 Observed stress for the initiation of plastic deformation for various 52100 steel structures

Structure	Hardness Rc	% Retained austenite	PD stress, psi
A	58	0	377,000
B	64	3.9	373,000
C	62	7.4	296,000
D	59	18.4	160,000

The location of the curves is a function of the retained austenite content; in general, the higher the retained austenite content the higher the rolling force for a given contact stress. The relationship between material structure and the stress at which plastic deformation was first noticed is also regular and is shown in Table 2. As the retained austenite content drops from 18.4 to 3.9 per cent the PD stress is increased from 160,000 to 373,000 psi. A further decrease of the retained austenite to zero does not substantially increase the resistance to initial plastic deformation. For the structures B, C, and D the stress necessary for the initiation of plastic deformation increases as the hardness number. While the hardness of structure A is considerably less than the hardness of structure B, the resistance to initial plastic deformation of the two structures remains practically the same. When structure A is compared with structure C it is evident that an increase in hardness is accompanied by a decrease in the stress necessary for the initiation of plastic deformation, a situation similar to that found with elastic limit data obtained in tension.

In Fig. 3 rolling friction results are compared for plates of structure A and C when run with a $^1/_2$-in-diam cemented-carbide ball. The change to the cemented-carbide ball, with its higher modulus of elasticity of 72×10^6 psi, alters the stress distribution in the region of the contact. In addition, the rolling force is measured for the contact pair cemented-carbide/steel instead of steel/steel as in Fig. 2. While the absolute values of the rolling force cannot be compared with those of Fig. 2, the relative values of the rolling force and the stress at which plastic deformation starts in the plates confirm the data obtained with steel balls.

The stress pattern is also affected by the diameter of the ball used in the run. In Fig. 4 the average rolling force is plotted against contact stress for runs using $^3/_4$-in-diam balls of the same structure as the plates. The higher rolling force observed for structure D than for structure C and the values of the stress necessary for plastic deformation both confirm the data that were obtained with the $^1/_2$-in-diam balls.

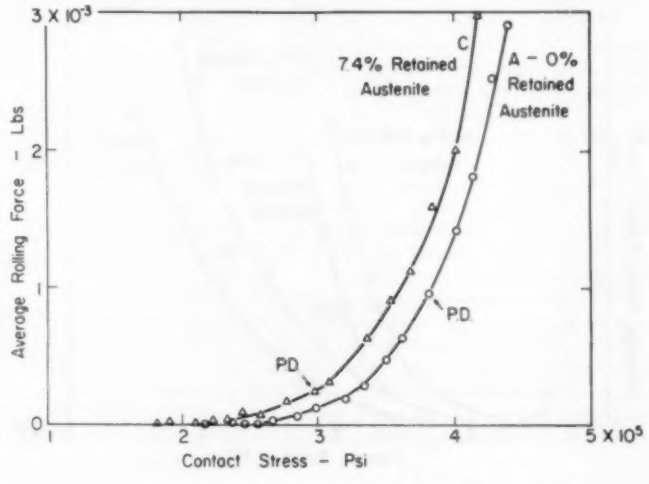


Fig. 3 Average rolling force versus contact stress for a 1/2-in-diam cemented-carbide ball rolling on plates of two different structures of SAE 52100 steel

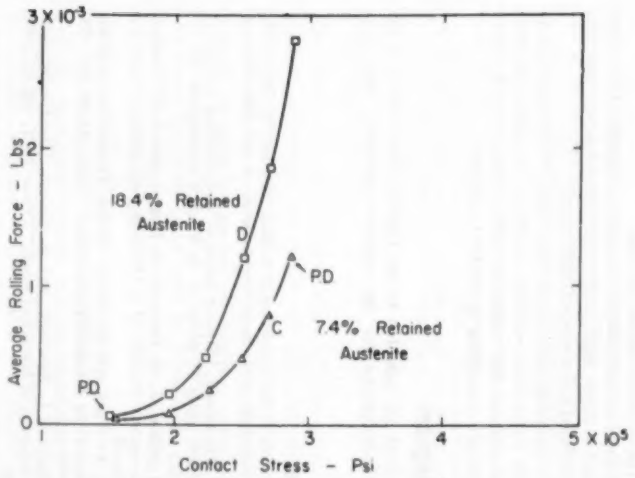


Fig. 4 Average rolling force versus contact stress for two different structures of SAE 52100 steel. Ball, 3/4-in. diam, of same structure as plates.

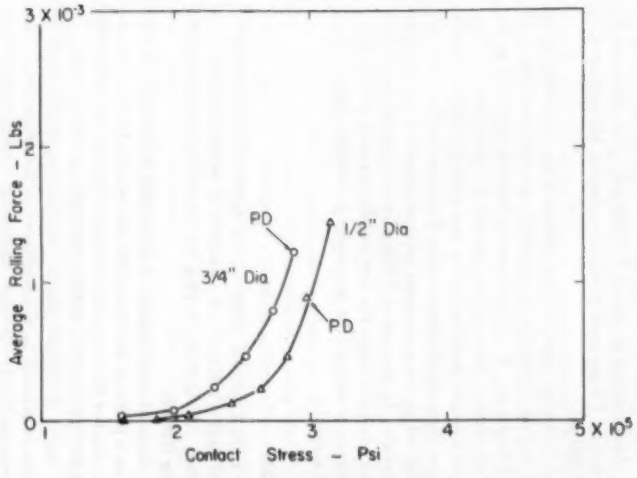


Fig. 5 Average rolling force versus contact stress for structure C, 7.4 per cent retained austenite, of SAE 52100 steel. Two different diameter balls of same structure as plates.

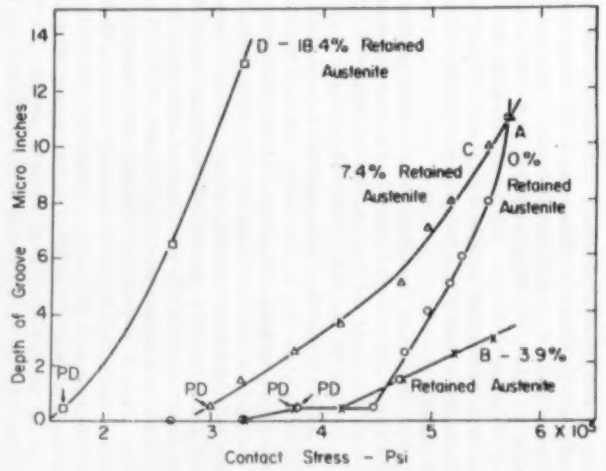


Fig. 6 Depth of groove developed versus contact stress for various structures of SAE 52100 steel plates. Cemented-carbide ball, 1/2-in. diam.

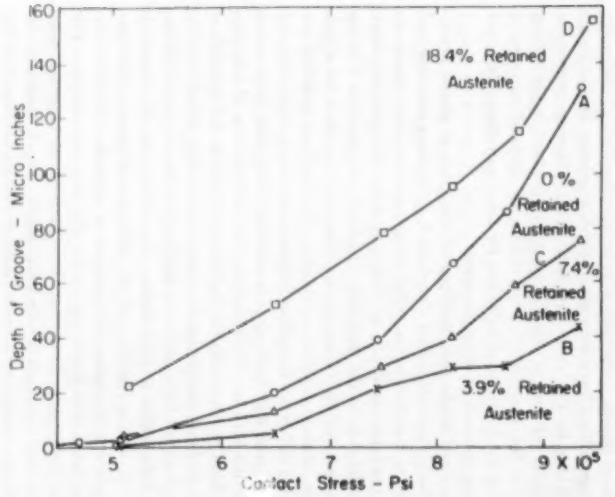


Fig. 7 Depth of groove developed versus contact stress for various structures of SAE 52100 steel plates. Cemented-carbide ball, 1/4-in. diam.

It is interesting to compare the rolling force versus contact stress curves for the same material structure but for different ball diameters. The curves for structure C and ball diameters 1/2 and 3/4 in. are shown in Fig. 5. While the stress necessary for the initiation of plastic deformation is independent of ball diameter, the average rolling force for a given stress is greater for the 3/4-in-diam ball than for the 1/2-in-diam ball. This difference represents the different volumes of stressed material involved in the rolling process for the two balls. The volume of stressed material is greater for the 3/4-in-diam ball than for the 1/2-in-diam ball even though the contact stress is the same.

On the basis of the rolling force and initiation of plastic deformation data it is evident that structure A, containing zero retained austenite, is superior to the other structures for a low stress instrument bearing application. Not only does this structure show a high resistance to the initiation of plastic deformation but the energy losses during rolling are considerably less than for structures that contain retained austenite. As has been pointed out before, this difference in rolling friction behavior is important for instrument bearings because it is essential that they have minimum friction torque. Fortunately, the combination of low friction and high resistance to the initiation of plastic deformation is achieved without sacrificing dimensional stability, which is a further requirement for a precision instrument bearing application.

Plastic Deformation at High Stress. In addition to the measurement of the rolling friction and the initiation of plastic deformation, the extent of plastic deformation at stresses well above the initiation point was investigated for the four structures. The depths of the grooves formed at these high contact stresses were measured with a surface profile instrument. Fig. 6 compares the groove depth formed by a 1/2-in-diam cemented-carbide ball for these structures at contact stresses up to approximately 600,000 psi. The points marked PD are again the stresses at which plastic deformation is first visually observed. In Fig. 7, the contact stress range is extended to almost 10^6 psi by using a 1/4-in-diam cemented-carbide ball on the same plate structures. At stresses up to 400,000 psi structures A and B exhibit the least amount of plastic deformation. Above about 470,000 psi structure B becomes superior, in its resistance to plastic deformation, to structure A. Above 570,000 psi structure A, Rc 58 becomes inferior to structure C, Rc 62. The superiority of the hardest structure, B, is maintained at these high contact stresses and the other structures resist plastic deformation in the order of their hardness numbers.

These observations help to explain the previously mentioned anomalous relationship between hardness and the stress necessary for the initiation of plastic deformation, that was obtained for the structures at the 58, 62, and 64 Rc levels. The rolling of a ball on a plane at stresses well in the plastic region of the material is a process of deformation akin to that which exists in a hardness test. In both cases excessive stresses are developed which severely deform the structure thus producing extensive work hardening and/or stress-induced phase transformations. At lower stresses, plastic deformation is just beginning and there is not enough deformation to promote extensive phase transformation to increase in turn the resistance to macroplastic deformation. At this initial stage, since retained austenite is relatively soft compared to tempered martensite, the effect of the austenite is to lower the resistance of the aggregate structure to microplastic deformation. Thus, as the data show, the extent of this weakening effect is a function of the amount of retained austenite present. On the basis of these data, up to 3.9 per cent retained austenite can be tolerated in the structure before a marked decrease in the resistance to microplastic deformation occurs. At high stresses, macroplastic deformation is extensive and is accompanied by the stress-induced transformation of retained austenite to martensite. In this stage the data show that the resistance to plastic deformation is a function of hardness. The more retained austenite present in the structure, at least up to a point, the greater is the contact stress necessary to complete the transformation to martensite. For structure B, which contained 3.9 per cent retained austenite, this occurred at about 470,000 psi. For structure C, which contained 7.4 per cent retained austenite, this occurred at about 570,000 psi.

A consideration of these findings indicates that hardness alone cannot be used as a basis for the selection of a heat-treatment for instrument bearings. In an instrument bearing application, the initiation of plastic deformation should be prevented; consequently, the superior material is the one with the highest elastic limit. From these results high elastic limit can be best achieved and maintained by using retained austenite content as the measuring yardstick. In the case of 52100 steel structures, those with no retained austenite have the best combination of elastic limit, rolling friction, resistance to microplastic deformation, and dimensional stability.

Conclusions

1 The rolling friction observed for hardened and tempered SAE 52100 steel is related to the amount of retained austenite in the structure. A structure having no retained austenite has the least rolling friction.

2 The stress at which plastic deformation is initiated, in rolling a ball on a plane, is also related to the retained austenite content. The smaller the retained austenite content the higher the contact stress necessary to initiate microplastic deformation.

3 The extent of contact plastic deformation which occurs at high stresses is related to the hardness of the structure. The hardest structure shows the least amount of macroplastic deformation.

4 Maximum hardness is not necessarily a good criterion for the selection of materials for instrument ball bearings that are designed to run at low contact stresses.

5 Retained austenite content is a superior yardstick on which to select and to insure high quality instrument bearing materials made of hardened and tempered 52100 steel.

6 In the case of hardened and tempered 52100 steel, maximum elastic limit, dimensional stability, resistance to plastic deformation at low stresses, and minimum rolling friction are achieved with a structure containing no retained austenite.

Acknowledgment

The authors are indebted to the Physics Department of the Research Laboratories for the retained austenite determinations.

References

1 D. Tabor, "The Mechanism of Rolling Friction—II, The Elastic Range," *Proceedings of the Royal Society*, series A, vol. 229, 1955, pp. 198–220.

2 R. C. Drutowski, "Energy Losses of Balls Rolling on Plates," TRANS. ASME, series D, JOURNAL OF BASIC ENGINEERING, vol. 81, 1959, pp. 233–238.

3 E. B. Mikus, J. M. Gerty, T. J. Hughel, and A. C. Knudsen, "The Dimensional Stability of a Precision Ball Bearing Material." to be published.

4 B. L. Averbach and M. Cohen, "X-Ray Determination of Retained Austenite by Integrated Intensities," *Trans. AIME*, vol. 176, 1948, pp. 401–415.

5 H. Muir, B. L. Averbach, and M. Cohen, "The Elastic Limit and Yield Behavior of Hardened Steels," *Trans. ASM*, vol. 47, 1955, pp. 380–398.

6 S. Timoshenko and J. H. Goodier, "Theory of Elasticity," McGraw-Hill Book Company, Inc., New York, N. Y., 1951, second edition, p. 372.

7 R. C. Drutowski, "The Volume of Stressed Material Involved in the Rolling of a Ball," to be published.

DISCUSSION

Robert A. Bareiss[3]

The subject matter which the authors have taken for their current work is of considerable importance to the manufacturers and users of antifriction bearings. It is therefore necessary that all aspects of this problem be given rigorous consideration. One of the major points made in this paper is the contact stress which causes the initiation of plastic deformation in the sample plates. In order to clearly establish the effects of retained austenite on plastic deformation, other contributing factors should be identified and accounted for in the results of any tests. In this regard, the authors should consider the effects of residual stresses on the initiation of plastic deformation in their tests. It has been shown by a number of authors (see Bibliography at end of this discussion) that internal stresses do play a considerable part in plastic deformation of a ball on a flat plate. Tensile residual stresses lower the contact stress required for plastic deformation while compressive residual stresses tend to raise slightly this contact stress. The change in contact stress to initiate deformation seems to be proportional to the magnitude of the residual stress. Since the test specimens used by the authors have in some cases contained a considerable proportion of retained austenite, it seems reasonable to assume that residual stresses of significant magnitude may also be present.

Bibliography

1 R. A. Bareiss, "Determination of Residual Stresses by Indentation Methods," WADC Technical Report 54-615, Wright Air Development Center, December, 1954.

2 F. Hoenig, "Proposal for a New Process for the Detection and Approximate Determination of Self-Stresses," *Berg und Huttenmannisches Jahrbuch*, vol. 182, no. 2, 1934, pp. 70–72 (in German).

3 A. O. Dervishyan, E. Janssen, and W. C. Hurly, "The Stress-Hardness Relation," Office of Ordnance Research, Contract No. DA-04-495-ORD-310, Report 54-76, September, 1954.

4 G. Sines and R. Carlson, "Hardness Measurements for Determination of Residual Stresses," *Bulletin, American Society for Testing Materials*, no. 180, February, 1952, pp. 35–37.

5 L. Hausseguy and H. Martinod, "A New Non-Destructive Procedure for the Determination of Superficial Residual Stresses," *La Recherche Aeronautique*, no. 37, January–February, 1954, pp. 43–50 (in French).

6 P. Blain, "The Effect of Internal Stress on Hardness," *Sheet Metal Industries*, January, 1949, pp. 135–136.

7 J. Pomey, L. Abel, and F. Goutel, "Measurement of Residual Stresses on the Surface of Case Hardened and Tempered Steels," *Comptes Rendus, Academie des Sciences, Paris*, vol. 228, May, 1949, pp. 1565–1567 (in French).

D. G. Flom[4]

The authors of this paper should be complimented on their contribution to bearing design and on the care with which their experiments were conducted. The important conclusion that hardness alone is insufficient to account for energy losses in pure rolling at low stress is of general applicability. In our own laboratory we have found that this is true for polymeric materials and that elastic hysteresis losses must be considered. As a consequence of this result, temperature and speed of rolling are found to be very important factors.

In view of Mr. Drutowski's earlier work on energy losses for balls rolling on plates I have wondered if the authors have used their data to calculate specific damping capacities for the 52100 steel as a function of retained austenite. In a very cursory search through the literature, I have found that studies of specific damping capacities for steel by more conventional techniques have been somewhat meager. In such a study one might again expect to find temperature and frequency effects to be important. In the case of rolling it would also be interesting to know if, within the elastic range, the dependence of friction on load and on ball diameter agree with the theoretical predictions of Tabor,[5] Drutowski,[6] and Flom and Bueche.[7] The latter dependence could perhaps be determined most easily with data using cemented carbide balls so that one would need to consider the deformation of the plate structure only.

As the authors show in Fig. 7, the resistance to plastic deformation of the steels during rolling at high stress appear directly related to their hardness numbers, indicating that a stage well into the plastic region has been reached. However, *full* plasticity has apparently not been reached since contact stress is still increasing with increasing depth of groove. It would be interesting to see whether extensions of the curves in Fig. 7 would bend upward sharply at contact stresses corresponding to yield pressures obtained from hardness measurements on the respective steels. If such a result were found it would indicate that time-dependent effects were negligible for the rolling speeds used in these experiments.

I should like to ask one question about the experimental results. The rolling forces given in the figures are presented as average values. Since the retained austenite in the steels would presumably exist in the bulk matrix alongside relatively hard and brittle martensite it is conceivable that variations in rolling force could be detected, especially at low stress, as regions of austenite and martensite came under the rolling ball. Were such variations observed or were these effects, if present, averaged out? If variations in rolling force existed, a correlation could perhaps be found by relating the distance between friction peaks and the per cent of retained austenite.

Authors' Closure

The comment of Mr. Bareiss is a valid one and the authors appreciate its possible significance. In addition, it should be pointed out that the size and distribution of the residual microstresses should be considered in any detailed explanation of plastic deformation. In speaking of the effect of residual stress on fatigue strength, Dolan [1],[8] has pointed out that the size and distribution of the residual microstresses in a specimen are more important than the average value of the microstresses (the value commonly obtained in residual stress measurements). While such a detailed consideration would undoubtedly enhance the interpretation of the experimental results, the authors feel that it would not basically affect the explanation offered, i.e., the influence of retained austenite on the ability of the material to resist microplastic deformation as shown by its effect on the elastic limit.

In answer to Mr. Flom, specific damping capacities were not calculated as a function of retained austenite. For the materials used, the specific damping capacity for a given ball diameter at a given contact stress varies directly as the rolling force. In the purely elastic contact region, i.e., below a load sufficient to cause plastic deformation, the specific damping capacity for a steel ball rolling on a steel plate is given by

$$\mathrm{SDC} = \frac{F}{3.84 \times 10^{-6} a^5 q_0^2}$$

obtained from equation (7) of a previous paper [2]. For a $^1/_2$-in-diam ball at a contact stress of 260,000 psi the specific

---

[3] Vice President, Lessells and Associates, Inc., Boston, Mass. Mem. ASME.

[4] General Electric Company, Schenectady, N. Y.

[5] D. Tabor, authors' Reference [1].

[6] R. C. Drutowski, authors' Reference [2].

[7] D. G. Flom and A. M. Bueche, "Theory of Rolling Friction for Spheres," *Journal of Applied Physics*, vol. 30, 1959, p. 1725.

[8] Numbers in brackets designate References at end of closure.

damping capacities of the zero and the 7.4 per cent retained austenite structures are, respectively, 0.06 and 0.22 per cent. These values are twice as large as would be obtained using Tabor's [3] original expression for elastic work in rolling but agree to within 3 per cent of his modified expression [4].

Tabor [3] observed experimentally that the rolling force increases as a n power of the load, the exponent being between 1.7 and 1.85. From theoretical considerations he predicted that at small strains the rolling force might be expected to increase as the $^4/_3$-power of the load. Flom and Beuche [5], using a different theoretical approach, also indicated that under certain conditions the rolling force should vary as the $^4/_3$-power of the load. Previously reported experimental data of Drutowski [2] agreed with these observations; for a cemented-carbide ball rolling on a hardened 52100 steel plate the rolling force was found to vary as the 1.2-power of the load at low loads and as the 2.4-power of the load at high loads. The data for the steel structures reported in this paper disagree with the theoretical predictions. For the data of Fig. 2 at the high contact stresses the rolling force varies as the 2.3-power of the load for the 18.4 per cent retained austenite structure, as the 2.1-power of the load for the 7.4 per cent retained austenite structure, and as the 1.9-power of the load for the 3.9 per cent and zero retained austenite structures. For the range of loads at the lowest contact stresses the exponents increase to 3.3 for the 7.4 per cent retained austenite structure, to 2.1 for the 3.9 per cent retained austenite structure, and to 2.9 for the zero retained austenite structure, and this increase is in opposition to the theoretical expectation. When a cemented-carbide ball rolls on two different structures at high contact stresses, as shown by the data of Fig. 3, the rolling force is found to vary as the 2.2-power of the load for the 7.4 per cent retained austenite structure and as the 2.7-power of the load for the zero retained austenite structure. For the lower contact stresses the exponents decrease, respectively, to 2.0 and 1.8. The deformation of the cemented-carbide ball is approximately 40 per cent that of the steel plates and should not be neglected; in fact, it appears that the contribution from the cemented carbide is enough to overpower the effect noted above for the steel/steel combinations.

The authors must apologize for not making clear in this presentation that the measurements are plotted as a function of nominal contact stress, a practice often employed in rolling contact literature. The nominal contact stress is that calculated from Hertz's formulas assuming no plastic deformation. If plastic deformation does occur the true value of the contact stress will be less than the nominal contact stress. For loads below that necessary to cause initiation of plastic deformation the true contact stress is equal to the nominal contact stress.

During the experiments the rolling force was continuously recorded and the variations of the rolling force along the rolling path noted. The presence of rolling friction peaks for even the smoothest steel surfaces has been previously reported [2]. For the specimens used in this investigation there appeared to be no obvious relationship between number and/or size of the peaks and the per cent of retained austenite, but no detailed check was made of a possible correlation.

References

1 T. J. Dolan, "Residual Stress, Strain Hardening and Fatigue," pp. 284–309 of the book: "Internal Stresses and Fatigue in Metals," edited by G. M. Rassweiler and W. L. Grube, Elsevier Publishing Company, 1959, sole distributor, D. Van Nostrand Company, Inc., New York, N. Y.

2 R. C. Drutowski, authors' reference [2].

3 D. Tabor, authors' reference [1].

4 J. A. Greenwood and D. Tabor, "The Friction of Hard Sliders on Lubricated Rubber: The Importance of Deformation Losses," *Proceedings of the Physical Society*, vol. 71, 1958, pp. 989–1001.

5 D. C. Flom and A. M. Bueche, D. G. Flom's reference [6].

A General Theory for Elastically Constrained Ball and Radial Roller Bearings Under Arbitrary Load and Speed Conditions

A. B. JONES
Chief Research Engineer
The Fafnir Bearing Company
New Britain, Conn.

A completely general solution is obtained whereby the elastic compliances of a system of any number of ball and radial roller bearings under any system of loads can be determined. Elastic yielding of the shaft and supporting structure are considered as well as centrifugal and gyroscopic loading of the rolling elements under high-speed operation. The solution defines the loading and attitude of each rolling element in each bearing of the system as well as the displacement of each inner ring with respect to its outer ring. For ball bearings the precise location of the load paths in each raceway are found. Life estimates can be more accurately made since the fatigue effects can be evaluated over known paths in the raceways. The solution, which is accomplished numerically by iterative techniques, has been programmed for an IBM-704 digital computer.

DESPITE their mechanical simplicity, rolling-element bearings, and in particular ball bearings, present some very complex problems when their actions under service conditions are studied.

All calculations of rolling element bearing fatigue life require a knowledge of the bearing's internal load distribution. Closed-form solutions for the forces on the rolling elements of a bearing are known for relatively few cases of loading. Existing closed-form solutions deal only with the case of parallel ring displacement and do not properly account for the change in contact angle found with ball bearings. Centrifugal ball loading and gyroscopic moments, important in high-speed bearings, are usually neglected.

The precise determination of elastic yields, of extreme importance of the designer of precision inertial guidance system, is not possible with existing closed-form solutions.

The reason for the inadequacy of existing solutions to properly define a bearing's internal load distribution lies in the complexity of the associated equations. Rolling element bearings constitute a statically-indeterminate, nonlinear, elastic system and the associated equations are not easily reducible to closed-form solutions without certain, simplifying assumptions.

The general equations for the elastic equilibrium of a ball bearing in three of the five possible degrees of freedom were given by the writer in Reference [1].[1] All known, closed-form solutions for different cases of loading can be deduced from these equations by making the necessary, simplifying assumptions. However, precise solutions could only be obtained by tedious, graphical or numerical integration of the pertinent equations.

With the advent of the modern, high-speed, digital computer, it became possible to quickly solve the most complex problem in bearing load distribution by numerical means.

This paper extends the scope of the original theory of Reference [1] so that an entire assemblage of bearings, shaft, and supporting structure is considered as a single, elastic system. The solution defines the elastic compliance of a point on the shaft with respect to the supporting structure in five degrees of freedom. Elastic yielding of the shaft and supporting structure are considered

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual statements of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 5, 1959. ASME Paper No. 59—Lub-10.

[1] Numbers in brackets designate References at end of paper.

Nomenclature

A = a fatigue constant for ball bearings
A' = a fatigue constant for roller bearings
B = total curvature factor $= f_o + f_i - 1$
$(CF)_q$ = centrifugal force on rolling element, lb
E = pitch diameter, in.
F_x, F_y, F_z = forces parallel to X, Y, and Z, lb
K_o, K_i = outer and inner-race deflection constants for ball bearings, in.$^{1/2} \times$ lb$^{-2/3}$
K_o', K_i' = outer and inner-race deflection constants for roller bearings; units depend on roller design
L = axial distance from point B on shaft to center of bearing inner raceway or to point of force application; L is positive when measured left from point B, in.
$_BL$ = fatigue life for ball bearing, hr
$_RL$ = fatigue life for roller bearing, hr
M_y, M_z = moments about axes parallel to Y and Z, lb in.
Mg_q = gyroscopic moment acting on ball, lb in.
N = speed of rotating race, rpm
P_D = mounted diametral clearance in roller bearing, in.
P_{oq}, P_{iq} = dynamic rolling-element loads at outer and inner-race contacts, lb
P_q = static rolling-element load, lb
Q_{oq}, Q_{iq}' = capacity of a ball or roller race contact for 90 per cent probability of survival to 10^6 revolutions of inner race, lb
X, Y, Z = orthogonal right-handed coordinate system
a_{oq}, a_{iq} = semimajor axes of outer and inner-race pressure ellipses, in.
b_{oq}, b_{iq} = semiminor axes of outer and inner-race pressure ellipses, in.
d = ball diameter, in.
f_o, f_i = ratio of transverse radius of ball race to ball diameter
j = total number of bearings and points at which shart is loaded by external or body forces

through the use of influence coefficients. The exact, internal load distribution is found for all of the bearings in the system. Centrifugal forces acting on the rolling elements and gyroscopic moments acting on the balls are considered. Bearing lives are evaluated by summation of the fatigue effects of the rolling elements' passages over precisely determined paths in each bearing raceway.

The present work has been programmed for an IBM-704 digital computer and, in addition to the elastic yields of the bearings and structural parts, and the bearings' fatigue lives, yields other pertinent data. Among these are contact area dimensions and stresses at each ball position, locations of the pressure-ellipse extremities, centrifugal force and gyroscopic moment at each rolling element position, theoretical fatigue life, frictional effects due to rolling, and many other items of interest.

Numerous, special-purpose programs for specific problems have also been developed using the basic theory outlined here. In several of these, the elastic deformations of the bearing rings, themselves, and their effect upon fatigue life have been considered.

The author wishes to acknowledge the assistance of Mr. George M. Cooley of The Fafnir Bearing Company in bringing the theory to its ultimate form.

Theoretical Development

Elastic Compliance of the System. Consider a system of p angular-contact ball bearings and t radial roller bearings mounted on a nonrigid shaft and supported in a nonrigid structure.

Let there be w points along the shaft where external loads or body forces are applied.

Then the total number of points on the shaft where loads or bearing reactions occur is $j = p + t + w$. The number of points at which the supporting structure is loaded is $u = p + t$.

Forces and moments and bearing reactions acting on the shaft can have components in accordance with the co-ordinate system of Fig. 1.

At each bearing position there will be structural yields and bearing deflections. These are referred to a similar co-ordinate system, Fig. 2.

Fig. 3 shows a portion of the shaft and supporting structure in their deflected state in XZ projection. A similar situation exists in XY projection.

In Fig. 3, u is any bearing location, B is a particular point on the shaft reference line. Shaft deflections and slopes are referred

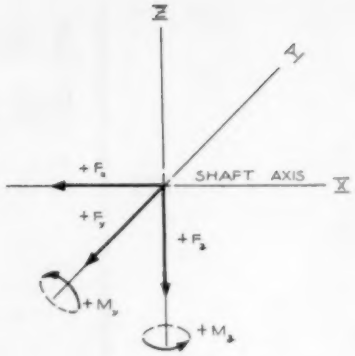


Fig. 1

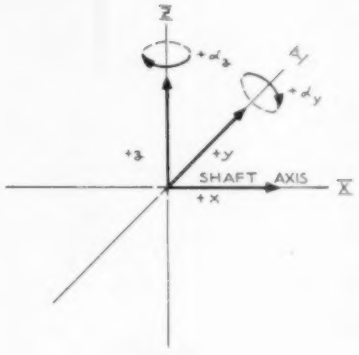


Fig. 2

to this reference line through influence coefficients. Deflections and slopes of the supporting structure are referred to the initial axis of the supporting structure through influence coefficients. $\bar{z}_u$ is the compliance of the inner ring of the bearing at u with respect to its outer ring in the Z-direction.

The shaft deflection at any bearing position u due to forces and moments in the XZ-plane is, relative to the shaft reference line

$$\Delta_{z_u} = [\Delta_{zuj}]\{F_{zj}\} + [\Delta'_{zuj}]\{M_{yj}\} \quad (1)$$

Nomenclature

l = length of roller, in.
p = number of ball bearings in system
q = rolling-element position index
= subscript indicating qth rolling-element position
r = radius of locus of ball-bearing inner-race curvature centers, in.
t = number of roller bearings in system
u = total number of bearings in system
w = number of points at which shaft is loaded by external or body forces
$\bar{x}, \bar{y}, \bar{z}$ = linear displacements parallel to X, Y, Z
α_q = angle between vector of ball's angular velocity about its own center and X-axis, radians
$\bar{\alpha}_y, \bar{\alpha}_z$ = angular displacements parallel to Y, Z, radians
β = initial contact angle of ball bearing after mounting
β_{oq}, β_{iq} = dynamic outer and inner-race operating contact angles
β_q = static operating contact angle
γ = ratio d/E
δ_{oq}, δ_{iq} = approach of rolling element to outer and inner races, in.
$\vartheta_{y_B}, \vartheta_{z_B}$ = angular displacements of shaft at point B about axes parallel to Y, and Z, radians
ω_{B_q} = angular velocity of ball about own center, rad × sec^{-1}
$\omega_{s_{oq}}, \omega_{s_{iq}}$ = angular velocity of spin of ball with respect to outer or inner race, rad × sec^{-1}
$\omega_{R_{oq}}, \omega_{R_{iq}}$ = angular velocity of rolling of ball with respect to outer or inner race, rad × sec^{-1}
Ω = angular velocity of rotating race, rad × sec^{-1}
Ω_{R_q} = orbital angular velocity of rolling element, rad × sec^{-1}

Subscripts $_o$ and $_i$ refer to conditions at outer and inner race contacts, respectively.

Subscript $_q$ refers to conditions at the qth rolling-element position.

Other notations as defined in text.

$\{F_{x_i}\}$ and $\{M_{y_i}\}$ are columns of forces and moments and are of order $j = p + t + w$.

$[\Delta_{xw_i}]$ and $[\Delta'_{xw_i}]$ are square matrices of influence coefficients, also of order j.

The slope of the shaft in the XZ-plane relative to the shaft reference line is

$$\alpha_{yu} = [\alpha_{yw_i}]\{F_{x_i}\} + [\alpha'_{yw_i}]\{M_{y_i}\} \tag{2}$$

The yield of the outer-race structure relative to the initial axis in the XZ-plane is

$$\nabla_{xu} = [\nabla_{xw_i}]\{F_{x_i}\} + [\nabla'_{xw_i}]\{M_{y_i}\} \tag{3}$$

and the slope of the outer-race supports relative to the initial axis is, in the XZ-plane

$$\rho_{yu} = [\rho_{yw_i}]\{F_{x_i}\} + [\rho'_{yw_i}]\{M_{y_i}\} \tag{4}$$

In the XY-plane a similar situation exists, so that

$$\Delta_{yu} = [\Delta_{yw_i}]\{F_{y_i}\} + [\Delta'_{yw_i}]\{M_{x_i}\} \tag{5}$$

$$\alpha_{xu} = [\alpha_{xw_i}]\{F_{y_i}\} + [\alpha'_{xw_i}]\{M_{x_i}\} \tag{6}$$

$$\nabla_{yu} = [\nabla_{yw_i}]\{F_{y_i}\} + [\nabla'_{yw_i}]\{M_{x_i}\} \tag{7}$$

$$\rho_{xu} = [\rho_{xw_i}]\{F_{y_i}\} + [\rho'_{xw_i}]\{M_{x_i}\} \tag{8}$$

The interest is in the sum of the shaft and support structure yields so that

$${}_1\Delta_{yu} = \Delta_{yu} + \nabla_{yu} = [{}_1\Delta_{yw_i}]\{F_{x_i}\} + [{}_1\Delta'_{yw_i}]\{M_{y_i}\} \tag{9}$$

$${}_1\alpha_{yu} = \alpha_{yu} + \rho_{yu} = [{}_1\alpha_{yw_i}]\{F_{x_i}\} + [{}_1\alpha'_{yw_i}]\{M_{y_i}\} \tag{10}$$

$${}_1\Delta_{yu} = \Delta_{yu} + \nabla_{yu} = [{}_1\Delta_{yw_i}]\{F_{y_i}\} + [{}_1\Delta'_{yw_i}]\{M_{x_i}\} \tag{11}$$

$${}_1\alpha_{xu} = \alpha_{xu} + \rho_{xu} = [{}_1\alpha_{xw_i}]\{F_{y_i}\} + [{}_1\alpha'_{xw_i}]\{M_{x_i}\} \tag{12}$$

where

$$[{}_1\Delta_{xw_i}] = [\Delta_{xw_i}] + [\nabla_{xw_i}] \text{ and so forth.} \tag{13}$$

Similarly, in the direction of the X-axis

$${}_1\Delta_{xu} = [{}_1\Delta_{xw_i}]\{F_{x_i}\} \tag{14}$$

where

$$[{}_1\Delta_{xw_i}] = [\Delta_{xw_i}] + [\nabla_{xw_i}] \tag{15}$$

If there are initial preloads which are reactive on elastic structural parts, then the columns of forces and moments must be adjusted accordingly.

$$\{F'_{x_i}\} = \{F_{x_i} - F'_{x_{i-w}}\} \tag{16}$$

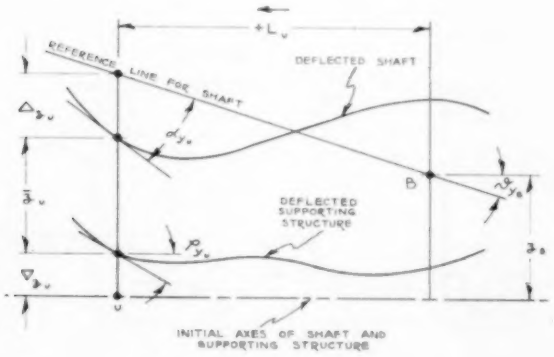


Fig. 3

$$\{F'_{y_i}\} = \{F_{y_i} - F'_{y_{i-w}}\} \tag{17}$$

$$\{F'_{z_i}\} = \{F_{z_i} - F'_{z_{i-w}}\} \tag{18}$$

$$\{M'_{y_i}\} = \{M_{y_i} - M'_{y_{i-w}}\} \tag{19}$$

$$\{M'_{z_i}\} = \{M_{z_i} - M'_{z_{i-w}}\} \tag{20}$$

The items $F'_{x_{i-w}}$ and so forth, are initial loads or moments which exist at a bearing position u when the system is at rest and before the application of external load.

If x_B, y_B, z_B, ϑ_{y_B}, and ϑ_{z_B} are displacements of the shaft reference line at point B relative to its initial position and attitude, the relative displacements of an inner race with respect to its outer race are

$$\bar{x}_u = x_B + \Delta''_{xu} + [{}_1\Delta_{xw_i}]\{F'_{x_i}\} \tag{21}$$

$$\bar{y}_u = y_B + \Delta''_{yu} + L_u\vartheta_{z_B} + [{}_1\Delta_{yw_i}]\{F'_{y_i}\} + [{}_1\Delta'_{yw_i}]\{M'_{z_i}\} \tag{22}$$

$$\bar{z}_u = z_B + \Delta''_{zu} + L_u\vartheta_{y_B} + [{}_1\Delta_{zw_i}]\{F'_{z_i}\} + [{}_1\Delta'_{zw_i}]\{M'_{y_i}\} \tag{23}$$

$$\bar{\alpha}_{yu} = \vartheta_{y_B} + \alpha''_{yu} + [{}_1\alpha_{yw_i}]\{F'_{z_i}\} + [{}_1\alpha'_{yw_i}]\{M'_{y_i}\} \tag{24}$$

$$\bar{\alpha}_{zu} = \vartheta_{z_B} + \alpha''_{zu} + [{}_1\alpha_{zw_i}]\{F'_{y_i}\} + [{}_1\alpha'_{zw_i}]\{M'_{z_i}\} \tag{25}$$

The items Δ''_{xw} and so forth, are initial deflections within the bearings as the result of preloading or the forcible application of initial constraints.

Five, nonlinear, simultaneous equations define the equilibrium of the system.

$$\sum_w F_{xw} + \sum_u F_{xu} = 0 = \psi_1 \tag{26}$$

$$\sum_w F_{yw} + \sum_u F_{yu} = 0 = \psi_2 \tag{27}$$

$$\sum_w F_{zw} + \sum_u F_{zu} = 0 = \psi_3 \tag{28}$$

$$\sum_w M_{yw} + \sum_w L_w F_{zw} + \sum_u M_{yu} + \sum_u L_u F_{zu} = 0 = \psi_4 \tag{29}$$

$$\sum_w M_{zw} + \sum_w L_w F_{yw} + \sum_u M_{zu} + \sum_u L_u F_{yu} = 0 = \psi_5 \tag{30}$$

These equations can be linearized and solved iteratively by the Newton-Raphson method.

Let $\delta_m{}^\circ$ be initial estimates of the variables x_B, y_B, z_B, ϑ_{y_B}, and ϑ_{z_B} in that sequence in Equations (26) through (30). Improved values are δ'_m.

$$\delta'_m = \delta_m{}^\circ - [a_{lm}]^{-1}\{\psi_l\} \tag{31}$$

where ψ_l are the error functions.

The elements a_{lm} of the matrix are

$$a_{11} = \sum_u \frac{\partial F_{xu}}{\partial \bar{x}_u} \tag{32}$$

$$a_{12} = \sum_u \frac{\partial F_{xu}}{\partial \bar{y}_u} \tag{33}$$

$$a_{13} = \sum_u \frac{\partial F_{xu}}{\partial \bar{z}_u} \tag{34}$$

$$a_{14} = \sum_u \left[L_u \frac{\partial F_{xu}}{\partial \bar{z}_u} + \frac{\partial F_{xu}}{\partial \bar{\alpha}_{yu}} \right] \tag{35}$$

$$a_{15} = \sum_u \left[L_u \frac{\partial F_{xu}}{\partial \bar{y}_u} + \frac{\partial F_{xu}}{\partial \bar{\alpha}_{zu}} \right] \tag{36}$$

$$a_{21} = \sum_u \frac{\partial F_{yu}}{\partial \bar{x}_u} \tag{37}$$

$$a_{21} = \sum_u \frac{\partial F_{yu}}{\partial \bar{y}_u} \tag{38}$$

$$a_{22} = \sum_u \frac{\partial F_{yu}}{\partial \bar{z}_u} \tag{39}$$

$$a_{24} = \sum_u \left[L_u \frac{\partial F_{yu}}{\partial \bar{z}_u} + \frac{\partial F_{yu}}{\partial \bar{\alpha}_{yu}} \right] \tag{40}$$

$$a_{25} = \sum_u \left[L_u \frac{\partial F_{yu}}{\partial \bar{y}_u} + \frac{\partial F_{yu}}{\partial \bar{\alpha}_{zu}} \right] \tag{41}$$

$$a_{31} = \sum_u \frac{\partial F_{zu}}{\partial \bar{x}_u} \tag{42}$$

$$a_{32} = \sum_u \frac{\partial F_{zu}}{\partial y_u} \tag{43}$$

$$a_{33} = \sum_u \frac{\partial F_{zu}}{\partial \bar{z}_u} \tag{44}$$

$$a_{34} = \sum_u \left[L_u \frac{\partial F_{zu}}{\partial \bar{z}_u} + \frac{\partial F_{zu}}{\partial \bar{\alpha}_{yu}} \right] \tag{45}$$

$$a_{35} = \sum_u \left[L_u \frac{\partial F_{zu}}{\partial \bar{y}_u} + \frac{\partial F_{zu}}{\partial \bar{\alpha}_{zu}} \right] \tag{46}$$

$$a_{41} = \sum_u \left[L_u \frac{\partial F_{zu}}{\partial \bar{x}_u} + \frac{\partial M_{yu}}{\partial \bar{x}_u} \right] \tag{47}$$

$$a_{42} = \sum_u \left[L_u \frac{\partial F_{zu}}{\partial \bar{y}_u} + \frac{\partial M_{yu}}{\partial \bar{y}_u} \right] \tag{48}$$

$$a_{43} = \sum_u \left[L_u \frac{\partial F_{zu}}{\partial \bar{z}_u} + \frac{\partial M_{yu}}{\partial \bar{z}_u} \right] \tag{49}$$

$$a_{44} = \sum_u \left(\left[L_u \frac{\partial F_{zu}}{\partial \bar{z}_u} + \frac{\partial M_{yu}}{\partial \bar{z}_u} \right] L_u + L_u \frac{\partial F_{zu}}{\partial \bar{\alpha}_{yu}} + \frac{\partial M_{yu}}{\partial \bar{\alpha}_{yu}} \right) \tag{50}$$

$$a_{45} = \sum_u \left(\left[L_u \frac{\partial F_{zu}}{\partial \bar{y}_u} + \frac{\partial M_{yu}}{\partial \bar{y}_u} \right] L_u + L_u \frac{\partial F_{zu}}{\partial \bar{\alpha}_{zu}} + \frac{\partial M_{yu}}{\partial \bar{\alpha}_{zu}} \right) \tag{51}$$

$$a_{51} = \sum_u \left[L_u \frac{\partial F_{yu}}{\partial \bar{x}_u} + \frac{\partial M_{zu}}{\partial \bar{x}_u} \right] \tag{52}$$

$$a_{52} = \sum_u \left[L_u \frac{\partial F_{yu}}{\partial \bar{y}_u} + \frac{\partial M_{zu}}{\partial \bar{y}_u} \right] \tag{53}$$

$$a_{53} = \sum_u \left[L_u \frac{\partial F_{yu}}{\partial \bar{z}_u} + \frac{\partial M_{zu}}{\partial \bar{z}_u} \right] \tag{54}$$

$$a_{54} = \sum_u \left(\left[L_u \frac{\partial F_{yu}}{\partial \bar{z}_u} + \frac{\partial M_{zu}}{\partial \bar{z}_u} \right] L_u + L_u \frac{\partial F_{yu}}{\partial \bar{\alpha}_{yu}} + \frac{\partial M_{zu}}{\partial \bar{\alpha}_{yu}} \right) \tag{55}$$

$$a_{55} = \sum_u \left(\left[L_u \frac{\partial F_{yu}}{\partial \bar{y}_u} + \frac{\partial M_{zu}}{\partial \bar{y}_u} \right] L_u + L_u \frac{\partial F_{yu}}{\partial \bar{\alpha}_{zu}} + \frac{\partial M_{zu}}{\partial \bar{\alpha}_{zu}} \right) \tag{56}$$

Forces and Moments in a Ball Bearing Under High-Speed Conditions. Fig. 4 shows an angular-contact ball bearing in its initial attitude and the five displacements the inner ring may have with respect to the outer.

Fig. 5 shows the convention of the ball position index q.

Fig. 6 shows the initial and final positions of the inner-race curvature center and ball center with respect to the outer-race curvature center at the qth ball position of a ball bearing at u.

$(AF)_q$ and $(RF)_q$ are axial and radial functions which relate the ring displacements to the ball forces and contact angles.

$$(AF)_q = Bd \sin \beta + \bar{x}_u + \bar{\alpha}_{yu} r_u \cos \frac{2\pi q}{n} + \bar{\alpha}_{zu} r_u \sin \frac{2\pi q}{n} \tag{57}$$

$$(RF)_q = Bd \cos \beta + \bar{y}_u \sin \frac{2\pi q}{n} + \bar{z}_u \cos \frac{2\pi q}{n} - \frac{\Delta P_D}{2} \tag{58}$$

where ΔP_D is any increase in diametral clearance of the mounted bearing.

If centrifugal forces and the friction forces arising from gyroscopic effects are considered, the forces and moments acting on the ball are shown in Fig. 7.

The ball loads are related to the normal approaches of the ball to the races through

$$P_{cq} = K_c \delta_{cq}^{3/2} \tag{59}$$

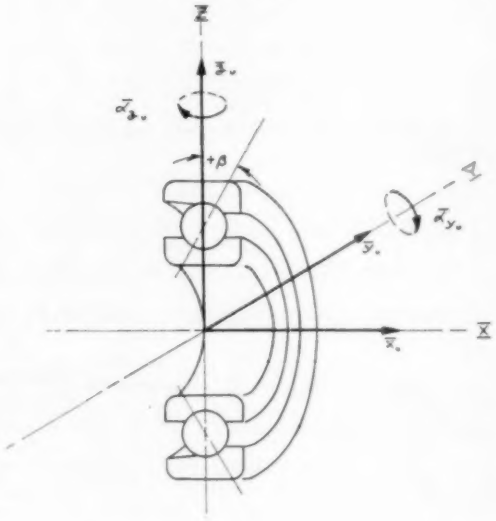


Fig. 4

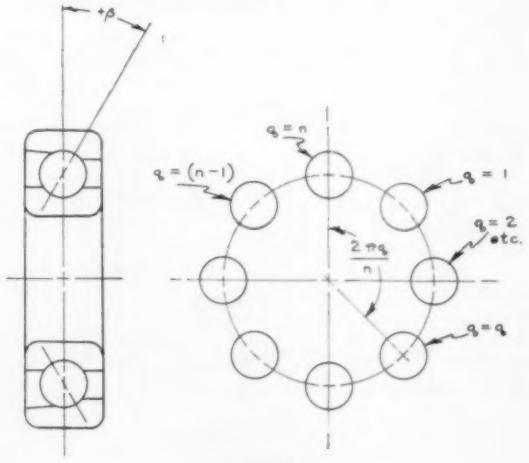


Fig. 5

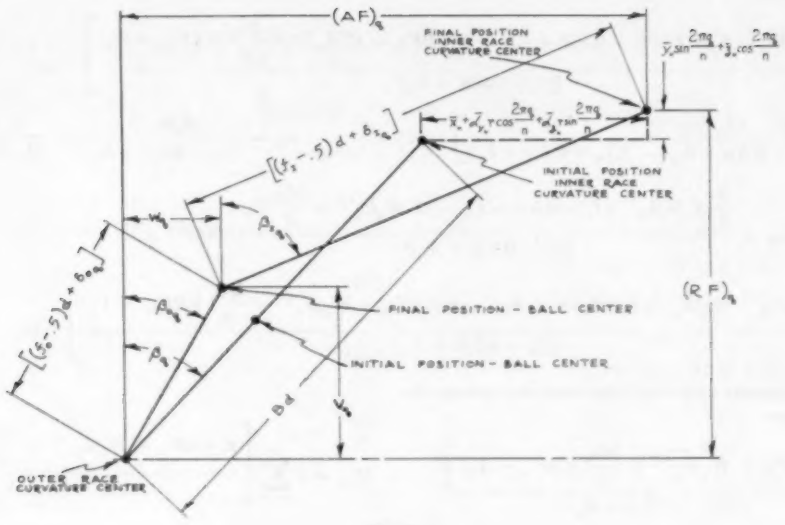

(AF)
FINAL POSITION INNER RACE CURVATURE CENTER
INITIAL POSITION INNER RACE CURVATURE CENTER
FINAL POSITION - BALL CENTER
INITIAL POSITION - BALL CENTER
(RF)
OUTER RACE CURVATURE CENTER

Fig. 6

and $$P_{i_q} = K_i \delta_{i_q}^{3/2} \tag{60}$$

At each ball position it is necessary to determine δ_{o_q}, δ_{i_q}, β_{o_q}, and β_{i_q}. Since the iterative techniques of the Newton-Raphson method will be used to solve the associated nonlinear equations the angles β_{o_q} and β_{i_q} are best stated in terms of the co-ordinates, V and W, in Fig. 6.

$$\sin \beta_{o_q} = \frac{W_q}{[(f_o - 0.5)d + \delta_{o_q}]} \tag{61}$$

$$\cos \beta_{o_q} = \frac{V_q}{[(f_o - 0.5)d + \delta_{o_q}]} \tag{62}$$

$$\sin \beta_{i_q} = \frac{(AF)_q - W_q}{[(f_i - 0.5)d + \delta_{i_q}]} \tag{63}$$

$$\cos \beta_{i_q} = \frac{(RF)_q - V_q}{[(f_i - 0.5)d + \delta_{i_q}]} \tag{64}$$

From Figs. 6 and 7 and Equations (61) through (64)

$$[(AF)_q - W_q]^2 + [(RF)_q - V_q]^2 - [(f_i - 0.5)d + \delta_{i_q}]^2 = 0 = \epsilon_1 \tag{65}$$

$$V_q^2 + W_q^2 - [(f_o - 0.5)d + \delta_{o_q}]^2 = 0 = \epsilon_2 \tag{66}$$

$$\frac{\left[K_o \delta_{o_q}^{3/2} V_q - \frac{M_{g_q} W_q}{d}\right]}{[(f_o - 0.5)d + \delta_{o_q}]} - \frac{\left[K_i \delta_{i_q}^{3/2}[(RF)_q - V_q] - \frac{M_{g_q}}{d}[(AF)_q - W_q]\right]}{[(f_i - 0.5)d + \delta_{i_q}]} - (CF)_q = 0 = \epsilon_3 \tag{67}$$

$$\frac{\left[K_o \delta_{o_q}^{3/2} W_q + \frac{M_{g_q} V_q}{d}\right]}{[(f_o - 0.5)d + \delta_{o_q}]} - \frac{\left[K_i \delta_{i_q}^{3/2}[(AF)_q - W_q] + \frac{M_{g_q}}{d}[(RF)_q - V_q]\right]}{[(f_i - 0.5)d + \delta_{i_q}]} = 0 = \epsilon_4 \tag{68}$$

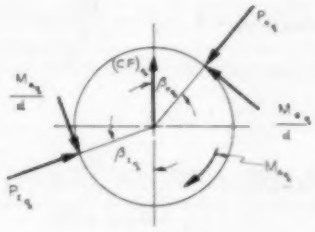


Fig. 7

If δ_k° are initial estimates of the variables V_q, W_q, δ_{o_q}, and δ_{i_q} in that order, improved values are δ_k', where

$$\delta_k' = \delta_k^\circ - [a_{ek}]^{-1}\{\epsilon_q\} \tag{69}$$

ϵ_q are the error functions, ϵ_1, ϵ_2, ϵ_3, and ϵ_4, in that order, from Equations (65) through (68).

The elements, a_{ek}, of the matrix are

$$a_{11} = -2[(RF)_q - V_q] \tag{70}$$

$$a_{12} = -2[(AF)_q - W_q] \tag{71}$$

$$a_{13} = 0 \tag{72}$$

$$a_{14} = -2[(f_i - 0.5)d + \delta_i] \tag{73}$$

$$a_{21} = 2V_q \tag{74}$$

$$a_{22} = 2W_q \tag{75}$$

$$a_{23} = -2[(f_o - 0.5)d + \delta_{o_q}] \tag{76}$$

$$a_{24} = 0 \tag{77}$$

$$a_{31} = \frac{K_o \delta_{o_q}^{3/2}}{[(f_o - 0.5)d + \delta_{o_q}]} + \frac{K_i \delta_{i_q}^{3/2}}{[(f_i - 0.5)d + \delta_{i_q}]} \tag{78}$$

$$a_{32} = -\frac{M_{g_q}}{d}\left[\frac{1}{[(f_o - 0.5)d + \delta_{o_q}]} + \frac{1}{[(f_i - 0.5)d + \delta_{i_q}]}\right] \tag{79}$$

$$a_{33} = \frac{\frac{3}{2} K_o V_q \delta_{o_q}^{1/2}[(f_o - 0.5)d + \delta_{o_q}] - K_o V_q \delta_{o_q}^{3/2} + \frac{M_{g_q}}{d} W_q}{[(f_o - 0.5)d + \delta_{o_q}]^2} \tag{80}$$

$$a_{34} = -\left[\frac{\frac{3}{2}K_i[(RF)_q - V_q]\delta_{i_q}^{1/2}[(f_i - 0.5)d + \delta_{i_q}] - K_i[(RF)_q - V_q]\delta_{i_q}^{3/2} + \frac{M_{G_q}}{d}[(AF)_q - W_q]}{[(f_i - 0.5)d + \delta_{i_q}]^2}\right] \quad (81)$$

$$a_{41} = \frac{M_{G_q}}{d}\left[\frac{1}{[(f_o - 0.5)d + \delta_{o_q}]} + \frac{1}{[(f_{i_q} - 0.5)d + \delta_{i_q}]}\right] \quad (82) \qquad a_{42} = \frac{K_o\delta_{o_q}^{3/2}}{[(f_o - 0.5)d + \delta_{o_q}]} + \frac{K_i\delta_{i_q}}{[(f_{i_q} - 0.5)d + \delta_{i_q}]} \quad (83)$$

$$a_{43} = \frac{\frac{3}{2}K_oW_q\delta_{o_q}^{1/2}[(f_o - 0.5)d + \delta_{o_q}] - K_oW_q\delta_{o_q}^{3/2} - \frac{M_{G_q}}{d}V_q}{[(f_o - 0.5)d + \delta_{o_q}]^2} \quad (84)$$

$$a_{44} = -\left[\frac{\frac{3}{2}K_i[(AF)_q - W_q]\delta_{i_q}^{1/2}[(f_i - 0.5)d + \delta_{i_q}] - K_i[(AF)_q - W_q]\delta_{i_q}^{3/2} - \frac{M_{G_q}}{d}[(RF)_q - V_q]}{[(f_i - 0.5)d + \delta_{i_q}]^2}\right] \quad (85)$$

The forces and moments with which the inner race acts on the shaft at its center are

$$F_{x_u} = \sum_{q=1}^{n}\left[\frac{K_i[(AF)_q - W_q]\delta_{i_q}^{3/2} + \frac{M_{G_q}}{d}[(RF)_q - V_q]}{[(f_i - 0.5)d + \delta_{i_q}]}\right] \quad (86)$$

$$F_{y_u} = \sum_{q=1}^{n}\left[\frac{K_i[(RF)_q - V_q]\delta_{i_q}^{3/2} - \frac{M_{G_q}}{d}[(AF)_q - W_q]}{[(f_i - 0.5)d + \delta_{i_q}]}\right]\sin\frac{2\pi q}{n} \quad (87)$$

$$F_{z_u} = \sum_{q=1}^{n}\left[\frac{K_i[(RF)_q - V_q]\delta_{i_q}^{3/2} - \frac{M_{G_q}}{d}[(AF)_q - W_q]}{[(f_i - 0.5)d + \delta_{i_q}]}\right]\cos\frac{2\pi q}{n} \quad (88)$$

$$M_{y_u} = r\sum_{q=1}^{n}\left[\frac{K_i[(AF)_q - W_q]\delta_{i_q}^{3/2} + \frac{M_{G_q}}{d}[(RF)_q - V_q]}{[(f_i - 0.5)d + \delta_{i_q}]} - \frac{f_i}{r_u}M_{G_q}\right]\cos\frac{2\pi q}{n} \quad (89)$$

$$M_{z_u} = r\sum_{q=1}^{n}\left[\frac{K_i[(AF)_q - W_q]\delta_{i_q}^{3/2} + \frac{M_{G_q}}{d}[(RF)_q - V_q]}{[(f_i - 0.5)d + \delta_{i_q}]} - \frac{f_i}{r_u}M_{G_q}\right]\sin\frac{2\pi q}{n} \quad (90)$$

For brevity let $\delta_u = (x_u, y_u, z_u, \alpha_{y_u}, \alpha_{z_u})$ as the case may be.

Then

$$\frac{\partial F_{z_u}}{\partial\delta_u} = \sum_{q=1}^{n}\frac{\left[\begin{array}{l}[(f_i - 0.5)d + \delta_{i_q}]\left\{K_i\delta_{i_q}^{3/2}\left(\frac{\partial(AF)_q}{\partial\delta_u} - \frac{\partial W_q}{\partial\delta_u}\right) + \frac{3}{2}K_i\delta_{i_q}^{1/2}[(AF)_q - W_q]\frac{\partial\delta_{i_q}}{\partial\delta_u}\right. \\ \left. + \frac{M_{G_q}}{d}\left(\frac{\partial(RF)_q}{\partial\delta_u} - \frac{\partial V_q}{\partial\delta_u}\right)\right\} - \left\{K_i\delta_{i_q}^{3/2}[(AF)_q - W_q] + \frac{M_{G_q}}{d}[(RF)_q - V_q]\right\}\frac{\partial\delta_{i_q}}{\partial\delta_u}\end{array}\right]}{[(f_i - 0.5)d + \delta_{i_q}]^2} \quad (91)$$

$$\frac{\partial F_{y_u}}{\partial\delta_u} = \sum_{q=1}^{n}\frac{\left[\begin{array}{l}[(f_i - 0.5)d + \delta_{i_q}]\left\{K_i\delta_{i_q}^{3/2}\left(\frac{\partial(RF)_q}{\partial\delta_u} - \frac{\partial V_q}{\partial\delta_u}\right) + \frac{3}{2}K_i\delta_{i_q}^{1/2}[(RF)_q - V_q]\frac{\partial\delta_{i_q}}{\partial\delta_u}\right. \\ \left. - \frac{M_{G_q}}{d}\left(\frac{\partial(AF)_q}{\partial\delta_u} - \frac{\partial W_q}{\partial\delta_u}\right)\right\} - \left\{K_i\delta_{i_q}^{3/2}[(RF)_q - V_q] - \frac{M_{G_q}}{d}[(AF)_q - W_q]\right\}\frac{\partial\delta_{i_q}}{\partial\delta_u}\end{array}\right]}{[(f_i - 0.5)d + \delta_{i_q}]^2}\sin\frac{2\pi q}{n} \quad (92)$$

$$\frac{\partial F_{z_u}}{\partial\delta_u} = \sum_{q=1}^{n}\frac{\left[\begin{array}{l}[(f_i - 0.5)d + \delta_{i_q}]\left\{K_i\delta_{i_q}^{3/2}\left(\frac{\partial(RF)_q}{\partial\delta_u} - \frac{\partial V_q}{\partial\delta_u}\right) + \frac{3}{2}K_i\delta_{i_q}^{1/2}[(RF)_q - V_q]\frac{\partial\delta_{i_q}}{\partial\delta_u}\right. \\ \left. - \frac{M_{G_q}}{d}\left(\frac{\partial(AF)_q}{\partial\delta_u} - \frac{\partial W_q}{\partial\delta_u}\right)\right\} - \left\{K_i\delta_{i_q}^{3/2}[(RF)_q - V_q] - \frac{M_{G_q}}{d}[(AF)_q - W_q]\right\}\frac{\partial\delta_{i_q}}{\partial\delta_u}\end{array}\right]}{[(f_i - 0.5)d + \delta_{i_q}]^2}\cos\frac{2\pi q}{n} \quad (93)$$

$$\frac{\partial M_{y_u}}{\partial\delta_u} = r\sum_{q=1}^{n}\frac{\left[\begin{array}{l}[(f_i - 0.5)d + \delta_{i_q}]\left\{K_i\delta_{i_q}^{3/2}\left(\frac{\partial(AF)_q}{\partial\delta_u} - \frac{\partial W_q}{\partial\delta_u}\right) + \frac{3}{2}K_i\delta_{i_q}^{1/2}[(AF)_q - W_q]\frac{\partial\delta_{i_q}}{\partial\delta_u}\right. \\ \left. + \frac{M_{G_q}}{d}\left(\frac{\partial(RF)_q}{\partial\delta_u} - \frac{\partial V_q}{\partial\delta_u}\right)\right\} - \left\{K_i\delta_{i_q}^{3/2}[(AF)_q - W_q] + \frac{M_{G_q}}{d}[(RF)_q - V_q]\right\}\frac{\partial\delta_{i_q}}{\partial\delta_u}\end{array}\right]}{[(f_i - 0.5)d + \delta_{i_q}]^2}\cos\frac{2\pi q}{n} \quad (94)$$

$$\frac{\partial M_{zu}}{\partial \delta_u} = r \sum_{q=1}^{n} \frac{\left[\begin{array}{l} [(f_i - 0.5)d + \delta_{i_q}] \left\{ K_i \delta_{i_q}^{3/2} \left(\dfrac{\partial (AF)_q}{\partial \delta_u} - \dfrac{\partial W_q}{\partial \delta_u} \right) + \dfrac{3}{2} K_i \delta_{i_q}^{1/2} [(AF)_q - W_q] \dfrac{\partial \delta_{i_q}}{\partial \delta_u} \right. \\ \left. + \dfrac{M_{G_q}}{d} \left(\dfrac{\partial (RF)_q}{\partial \delta_u} - \dfrac{\partial V_q}{\partial \delta_u} \right) \right\} - \left\{ K_i \delta_{i_q}^{3/2} [(AF)_q - W_q] + \dfrac{M_{G_q}}{d} [(RF)_q - V_q] \right\} \dfrac{\partial \delta_{i_q}}{\partial \delta_u} \end{array} \right]}{[(f_i - 0.5)d + \delta_{i_q}]^2} \sin \frac{2\pi q}{n} \quad (95)$$

The derivatives of $(AF)_q$ and $(RF)_q$ with respect to the δ_u are

$$\frac{\partial (AF)_q}{\partial x_u} = 1 \quad (96)$$

$$\frac{\partial (AF)_q}{\partial y_u} = 0 \quad (97)$$

$$\frac{\partial (AF)_q}{\partial z_u} = 0 \quad (98)$$

$$\frac{\partial (AF)_q}{\partial \bar{\alpha}_{y_u}} = r \cos \frac{2\pi q}{n} \quad (99)$$

$$\frac{\partial (AF)_q}{\partial \bar{\alpha}_{z_u}} = r \sin \frac{2\pi q}{n} \quad (100)$$

$$\frac{\partial (RF)_q}{\partial x_u} = 0 \quad (101)$$

$$\frac{\partial (RF)_q}{\partial y_u} = \sin \frac{2\pi q}{n} \quad (102)$$

$$\frac{\partial (RF)_q}{\partial z_u} = \cos \frac{2\pi q}{n} \quad (103)$$

$$\frac{\partial (RF)_q}{\partial \bar{\alpha}_{y_u}} = 0 \quad (104)$$

$$\frac{\partial (RF)_q}{\partial \bar{\alpha}_{z_u}} = 0 \quad (105)$$

If now Equations (65) through (68) are differentiated with respect to the δ_u, four simultaneous equations result which are linear in $\frac{\partial V_q}{\partial \delta_u}$, $\frac{\partial W_q}{\partial \delta_u}$, $\frac{\partial \delta_{c_q}}{\partial \delta_u}$, and $\frac{\partial \delta_{i_q}}{\partial \delta_u}$ and from which the latter can be found. The solutions are

$$\begin{Bmatrix} \frac{\partial V_q}{\partial \delta_u} \\ \frac{\partial W_q}{\partial \delta_u} \\ \frac{\partial \delta_{c_q}}{\partial \delta_u} \\ \frac{\partial \delta_{i_q}}{\partial \delta_u} \end{Bmatrix} = [b_{ij}]^{-1} \begin{Bmatrix} [(AF)_q - W_q] \frac{\partial (AF)_q}{\partial \delta_u} + [(RF)_q - V_q] \frac{\partial (RF)_q}{\partial \delta_u} \\ 0 \\ \frac{K_i \delta_{i_q}^{3/2} \frac{\partial (RF)_q}{\partial \delta_u} - \frac{M_{G_q}}{d} \frac{\partial (AF)_q}{\partial \delta_u}}{[(f_i - 0.5)d + \delta_{i_q}]} \\ \frac{K_i \delta_{i_q}^{3/2} \frac{\partial (AF)_q}{\partial \delta_u} + \frac{M_{G_q}}{d} \frac{\partial (RF)_q}{\partial \delta_u}}{[(f_i - 0.5)d + \delta_{i_q}]} \end{Bmatrix} \quad (106)$$

The elements, b_{ij}, of the matrix are

$$[b_{11} = (RF)_q - V_q] \quad (107)$$

$$b_{12} = [(AF)_q - W_q] \quad (108)$$

$$b_{13} = 0 \quad (109)$$

$$b_{14} = [(f_i - 0.5)d + \delta_{i_q}] \quad (110)$$

$$b_{21} = V_q \quad (111)$$

$$b_{22} = W_q \quad (112)$$

$$b_{23} = -[(f_o - 0.5)d + \delta_{o_q}] \quad (113)$$

$$b_{24} = 0 \quad (114)$$

The remaining elements b_{31}, b_{32}, b_{33}, b_{34}, b_{41}, b_{42}, b_{43}, and b_{44} are, respectively, identical to a_{31}, a_{32}, a_{33}, a_{34}, a_{41}, a_{42}, a_{43}, and a_{44}, Equations (78) through (85).

The centrifugal force and the gyroscopic moment are

$$(CF)_q = C_1 d^3 E \Omega^2 \left(\frac{\Omega_E}{\Omega} \right)_q^2 \quad (115)$$

$$M_{G_q} = C_2 d^5 \left(\frac{\omega_E}{\Omega} \right)_q \left(\frac{\Omega_E}{\Omega} \right)_q \Omega^2 \sin \alpha_q \quad (116)$$

C_1 and C_2 depend on material density. Ω is the angular velocity of the rotating race.

The ratios (Ω_E/Ω) and (ω_E/Ω) depend upon which race is rotating and upon the type of ball "control." In angular-contact ball bearings it is found that the ball will roll on one race without spin and that all spin occurs with respect to the other race. If no spin occurs at a race contact the ball is said to be controlled by that race. If gyroscopic slip of the ball is prevented, the ball's angular-velocity vector lies in the plane of the bearing's rotational axis and its attitude depends on the type of control.

For outer-race control

$$\tan \alpha_q = \frac{\sin \beta_{c_q}}{\cos \beta_{c_q} + \gamma} \quad (117)$$

For inner-race control

$$\tan \alpha_q = \frac{\sin \beta_{i_q}}{\cos \beta_{i_q} - \gamma} \quad (118)$$

Table 1 shows values of $\left(\frac{\Omega_E}{\Omega} \right)_q$.

The ratio $\left(\frac{\omega_E}{\Omega} \right)_q$ is

$$\left(\frac{\omega_E}{\Omega} \right)_q = \frac{\pm 1}{\left[\frac{\cos \beta_{c_q} + \tan \alpha_q \sin \beta_{c_q}}{1 + \gamma \cos \beta_{c_q}} + \frac{\cos \beta_{i_q} + \tan \alpha_q \sin \beta_{i_q}}{1 - \gamma \cos \beta_{i_q}} \right] \gamma \cos \alpha_q} \quad (119)$$

Table 1

	Value of $\left(\frac{\Omega_E}{\Omega}\right)_q$	
	Outer-race control	Inner-race control
Outer rotating	$\frac{\cos(\beta_{i_q} - \beta_{o_q}) + \gamma \cos \beta_{i_q}}{1 + \cos(\beta_{i_q} - \beta_{o_q})}$	$\frac{1 + \gamma \cos \beta_{o_q}}{1 + \cos(\beta_{i_q} - \beta_{o_q})}$
Inner rotating	$\frac{1 - \gamma \cos \beta_{i_q}}{1 + \cos(\beta_{i_q} - \beta_{o_q})}$	$\frac{\cos(\beta_{i_q} - \beta_{o_q}) - \gamma \cos \beta_{o_q}}{1 + \cos(\beta_{i_q} - \beta_{o_q})}$

The upper sign is used with outer-race rotation; the lower sign with inner-race rotation.

It is necessary to make an assumption as to the type of control. A check on the correctness of the assumption is given by the following relations obtained from consideration of the moment required to spin a ball relative to a raceway about the normal to the pressure area at its center.

Outer-race control exists at a ball position q if

$$P_{o_q} a_{o_q} E(\epsilon)_{o_q} \cos(\beta_{i_q} - \beta_{o_q}) > P_{i_q} a_{i_q} E(\epsilon)_{i_q} \tag{120}$$

Inner-race control requires that

$$P_{i_q} a_{i_q} E(\epsilon)_{i_q} \cos(\beta_{i_q} - \beta_{o_q}) > P_{o_q} a_{o_q} E(\epsilon)_{o_q} \tag{121}$$

$E(\epsilon)$ is the complete elliptic integral of the second kind having the modulus $\sin \epsilon$ where

$$\cos \epsilon = \frac{b}{a} \tag{122}$$

b and a are the semiaxes of the pressure ellipse at a race contact, calculated in accordance with the Hertz theory.

Under certain conditions of operation a ball may be completely out of contact with the inner race. At that time the outer-race contact angle β_{o_q} and the gyroscopic moment M_{G_q} are zero. The force P_{o_q} is equal to the centrifugal force.

Loss of inner-race contact will occur when

$$(AF)_q^2 + [(RF)_q - (f_o - 0.5)d - K_o^{-2/3}(CF)_q^{2/3}]^2 \leqslant [(f_i - 0.5)d]^2 \tag{123}$$

Forces and Moments in a Ball Bearing Under Static Loading. If the centrifugal force and the gyroscopic moment acting on a ball are negligible, the assumption $\Omega = 0$ in the foregoing will yield a solution for the case of static loading. However, the computations are unnecessarily long since many of the equations become redundant.

The solution for static loading can be obtained from consideration of the potential energy in a deflected bearing. Here, however, to preserve continuity the static case is deduced from the dynamic case as follows

With $\Omega = 0$:

$$P_{o_q} = P_{i_q} = P_q \tag{124}$$

$$\beta_{o_q} = \beta_{i_q} = \beta_q \tag{125}$$

and

$$\tan \beta_q = \frac{(AF)_q}{(RF)_q} \tag{126}$$

The total approach of inner and outer races at a ball position is

$$\delta_q = [(AF)^2 + (RF)^2]^{1/2} - Bd \geqslant 0 \tag{127}$$

The ball load is

$$P_q = K\delta_q^{3/2} \tag{128}$$

where

$$K = \left[\frac{1}{K_o^{3/2}} + \frac{1}{K_i^{3/2}}\right]^{-2/3} \tag{129}$$

The forces and moments with which the inner race acts on the shaft at its center are

$$F_{x_u} = \sum_{q=1}^{n} P_q \sin \beta_q \tag{130}$$

$$F_{y_u} = \sum_{q=1}^{n} P_q \cos \beta_q \sin \frac{2\pi q}{n} \tag{131}$$

$$F_{z_u} = \sum_{q=1}^{n} P_q \cos \beta_q \cos \frac{2\pi q}{n} \tag{132}$$

$$M_{y_u} = r \sum_{q=1}^{n} P_q \sin \beta_q \cos \frac{2\pi q}{n} \tag{133}$$

$$M_{z_u} = r \sum_{q=1}^{n} P_q \sin \beta_q \sin \frac{2\pi q}{n} \tag{134}$$

The associated derivatives are

$$\frac{\partial F_{x_u}}{\partial \bar{x}_u} = K \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \tag{135}$$

$$\frac{\partial F_{x_u}}{\partial \bar{y}_u} = K \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \tag{136}$$

$$\frac{\partial F_{x_u}}{\partial \bar{z}_u} = K \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \cos \frac{2\pi q}{n} \tag{137}$$

$$\frac{\partial F_{x_u}}{\partial \bar{\alpha}_{y_u}} = rK \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \cos \frac{2\pi q}{n} \tag{138}$$

$$\frac{\partial F_{x_u}}{\partial \bar{\alpha}_{z_u}} = rK \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \sin \frac{2\pi q}{n} \tag{139}$$

$$\frac{\partial F_{y_u}}{\partial \bar{x}_u} = K \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \tag{140}$$

$$\frac{\partial F_{y_u}}{\partial \bar{y}_u} = K \sum_{q=1}^{n} [U_q + V_q \cos^2 \beta_q] \sin^2 \left(\frac{2\pi q}{n}\right) \tag{141}$$

$$\frac{\partial F_{y_u}}{\partial \bar{z}_u} = K \sum_{q=1}^{n} [U_q + V_q \cos^2 \beta_q] \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \tag{142}$$

$$\frac{\partial F_{yz}}{\partial \bar{\alpha}_{yz}} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (143)$$

$$\frac{\partial F_{yz}}{\partial \bar{\alpha}_{xz}} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin^2 \left(\frac{2\pi q}{n}\right) \quad (144)$$

$$\frac{\partial F_{xz}}{\partial \bar{x}_z} = K \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \cos \frac{2\pi q}{n} \quad (145)$$

$$\frac{\partial F_{xz}}{\partial y_z} = K \sum_{q=1}^{n} [U_q + V_q \cos^2 \beta_q] \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (146)$$

$$\frac{\partial F_{xz}}{\partial \bar{z}_z} = K \sum_{q=1}^{n} [U_q + V_q \cos^2 \beta_q] \cos^2 \left(\frac{2\pi q}{n}\right) \quad (147)$$

$$\frac{\partial F_{xz}}{\partial \bar{\alpha}_{yz}} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \cos^2 \left(\frac{2\pi q}{n}\right) \quad (148)$$

$$\frac{\partial F_{xz}}{\partial \bar{\alpha}_{xz}} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (149)$$

$$\frac{\partial M_{yz}}{\partial \bar{x}_z} = rK \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \cos \frac{2\pi q}{n} \quad (150)$$

$$\frac{\partial M_{yz}}{\partial y_z} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (151)$$

$$\frac{\partial M_{yz}}{\partial \bar{z}_z} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \cos^2 \left(\frac{2\pi q}{n}\right) \quad (152)$$

$$\frac{\partial M_{yz}}{\partial \bar{\alpha}_{yz}} = r^2 K \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \cos^2 \left(\frac{2\pi q}{n}\right) \quad (153)$$

$$\frac{\partial M_{yz}}{\partial \bar{\alpha}_{xz}} = r^2 K \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (154)$$

$$\frac{\partial M_{xz}}{\partial \bar{x}_z} = rK \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \sin \frac{2\pi q}{n} \quad (155)$$

$$\frac{\partial M_{xz}}{\partial y_z} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin^2 \left(\frac{2\pi q}{n}\right) \quad (156)$$

$$\frac{\partial M_{xz}}{\partial \bar{z}_z} = rK \sum_{q=1}^{n} V_q \sin \beta_q \cos \beta_q \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (157)$$

$$\frac{\partial M_{xz}}{\partial \bar{\alpha}_{yz}} = r^2 K \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \sin \frac{2\pi q}{n} \cos \frac{2\pi q}{n} \quad (158)$$

$$\frac{\partial M_{xz}}{\partial \bar{\alpha}_{xz}} = r^2 K \sum_{q=1}^{n} [U_q + V_q \sin^2 \beta_q] \sin^2 \left(\frac{2\pi q}{n}\right) \quad (159)$$

where

$$U_q = \frac{\delta_q^{1/2}}{[\delta_q + (Bd)]} \quad (160)$$

$$V_q = \frac{\delta_q^{1/2}[\delta_q + 3(Bd)]}{2[\delta_q + (Bd)]} \quad (161)$$

Forces in a Radial Roller Bearing Under High-Speed Conditions. A radial roller bearing can be loaded only by the displacements y_z and z_z if misalignment effects are small.

For a radial roller bearing with initial diametral clearance P_D when mounted, the total compression at a rolling element is

$$\delta_q = y_z \sin \frac{2\pi q}{n} + z_z \cos \frac{2\pi q}{n} - \frac{P_D}{2} > 0 \quad (162)$$

δ_q must be positive since the roller cannot take tension.

There exists no precise formula for calculating the approach of a roller to a raceway as this depends somewhat on the nature and degree of the crowning of the roller or raceway. However, an expression of the form

$$P_{oq} = K_o' \delta_{oq}^{\ b} \quad (163)$$

$$P_{iq} = K_i' \delta_{iq}^{\ b} \quad (164)$$

can be fitted to experimental data.

The centrifugal force on a roller is

$$(CF)_q = c_9 d^3 l E \Omega^2 [1 \mp \gamma]^3 \quad (165)$$

The upper sign applies to inner-race rotation; the lower to outer.

The total compression of a rolling element is the sum of the compressions at the inner and outer-race contacts.

$$\delta_q = \delta_{oq} + \delta_{iq} \quad (166)$$

also

$$P_{oq} = P_{iq} + (CF)_q \quad (167)$$

and

$$K_o'(\delta_q - \delta_{iq})^b - K_i' \delta_{iq}^{\ b} - (CF)_q = 0 = \Lambda_q \quad (168)$$

if $(CF)_q \geqslant K_o' \delta_q^{\ b}$, $P_{oq} = (CF)_q$, and δ_{iq} and P_{iq} are zero.

If b is other than unity, establish the error function Λ_q. If δ_{iq}° is an estimate of δ_{iq}, a better estimate is δ_{iq}', where

$$\delta'_{iq} = \delta^{\circ}_{iq} + \frac{\Lambda_q}{b[K_o'(\delta_q - \delta_{iq})^{b-1} + K_i' \delta_{iq}^{\ b-1}]} > 0 \quad (169)$$

and so forth.

If b is unity

$$\delta_{iq} = \frac{[K_o' \delta_q - (CF)_q]}{[K_o' + K_i']} > 0 \quad (170)$$

$$P_{iq} = \frac{[K_o' \delta_q - (CF)_q]}{\left[\frac{K_o'}{K_i'} + 1\right]} > 0 \quad (171)$$

The forces with which the roller bearing acts on the shaft are

$$F_{xz} = 0 \quad (172)$$

$$F_{yz} = K_i' \sum_{q=1}^{n} \delta^b_{iq} \sin \frac{2\pi q}{n} \quad (173)$$

$$F_{zz} = K_i' \sum_{q=1}^{n} \delta^b_{iq} \cos \frac{2\pi q}{n} \quad (174)$$

$$M_{yz} = 0 \quad (175)$$

$$M_{zz} = 0 \quad (176)$$

The associated derivatives are

$$\frac{\partial F_{xz}}{\partial(\bar{x}_z, y_z, z_z, \bar{\alpha}_{yz}, \bar{\alpha}_{xz})} = 0 \quad (177)$$

$$\frac{\partial F_{yz}}{\partial(\bar{x}_z, \bar{\alpha}_{yz}, \bar{\alpha}_{xz})} = 0 \quad (178)$$

$$\frac{\partial F_{yu}}{\partial \bar{y}_u} = K'_i b \sum_{q=1}^{n} \frac{\delta^{b-1}{}_{i_q} \sin^2\left(\frac{2\pi q}{n}\right)}{1 + \frac{K'_i}{K'_o}\left[\frac{\delta_{i_q}}{\delta_q - \delta_{i_q}}\right]^{b-1}} \tag{179}$$

$$\frac{\partial F_{yu}}{\partial \bar{z}_u} = K'_i b \sum_{q=1}^{n} \frac{\delta^{b-1}{}_{i_q} \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n}}{1 + \frac{K'_i}{K'_o}\left[\frac{\delta_{i_q}}{\delta_q - \delta_{i_q}}\right]^{b-1}} \tag{180}$$

$$\frac{\partial F_{zu}}{\partial(\bar{x}_u, \bar{\alpha}_{yu}, \bar{\alpha}_{zu})} = 0 \tag{181}$$

$$\frac{\partial F_{zu}}{\partial \bar{y}_u} = K'_i b \sum_{q=1}^{n} \frac{\delta_{i_q}{}^{b-1} \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n}}{1 + \frac{K'_i}{K'_o}\left[\frac{\delta_{i_q}}{\delta_q - \delta_{i_q}}\right]^{b-1}} \tag{182}$$

$$\frac{\partial F_{zu}}{\partial \bar{z}_u} = K'_i b \sum_{q=1}^{n} \frac{\delta^{b-1}{}_{i_q} \cos^2\left(\frac{2\pi q}{n}\right)}{1 + \frac{K'_i}{K'_o}\left[\frac{\delta_{i_q}}{\delta_q - \delta_{i_q}}\right]^{b-1}} \tag{183}$$

$$\frac{\partial M_{yu}}{\partial(\bar{x}_u, \bar{y}_u, \bar{z}_u, \bar{\alpha}_{yu}, \bar{\alpha}_{zu})} = 0 \tag{184}$$

$$\frac{\partial M_{zu}}{\partial(\bar{x}_u, \bar{y}_u, \bar{z}_u, \bar{\alpha}_{yu}, \bar{\alpha}_{zu})} = 0 \tag{185}$$

If b is unity

$$F_{yu} = \frac{1}{\left[\frac{K'_o}{K'_i} + 1\right]} \sum_{q=1}^{n} [K'_o \delta_q - (CF)_q] \sin\frac{2\pi q}{n} \tag{186}$$

$$F_{zu} = \frac{1}{\left[\frac{K'_o}{K'_i} + 1\right]} \sum_{q=1}^{n} [K'_o \delta_q - (CF)_q] \cos\frac{2\pi q}{n} \tag{187}$$

$$\frac{\partial F_{yu}}{\partial \bar{y}_u} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \sin^2\left(\frac{2\pi q}{n}\right) \tag{188}$$

$$\frac{\partial F_{yu}}{\partial \bar{z}_u} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n} \tag{189}$$

$$\frac{\partial F_{zu}}{\partial \bar{y}_u} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n} \tag{190}$$

$$\frac{\partial F_{zu}}{\partial \bar{z}_u} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \cos^2\left(\frac{2\pi q}{n}\right) \tag{191}$$

Forces in a Radial Roller Bearing Under Static Load. Under static load

$$(CF)_q = 0 \tag{192}$$

$$P_{z_q} = P_{i_q} = P_q \tag{193}$$

$$\delta_{i_q} = \frac{\delta_q}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]} \tag{194}$$

$$F_{yu} = \frac{K'_i}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta_q{}^b \sin\frac{2\pi q}{n} \tag{195}$$

$$F_{zu} = \frac{K'_i}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta^b{}_q \cos\frac{2\pi q}{n} \tag{196}$$

$$\frac{\partial F_{yu}}{\partial \bar{y}_u} = \frac{K'_i b}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta^{b-1}{}_q \sin^2\left(\frac{2\pi q}{n}\right) \tag{197}$$

$$\frac{\partial F_{yu}}{\partial \bar{z}_u} = \frac{K'_i b}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta^{b-1}{}_q \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n} \tag{198}$$

$$\frac{\partial F_{zu}}{\partial \bar{y}_u} = \frac{K'_i b}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta^{b-1}{}_q \sin\frac{2\pi q}{n} \cos\frac{2\pi q}{n} \tag{199}$$

$$\frac{\partial F_{zu}}{\partial \bar{z}_u} = \frac{K'_i b}{\left[1 + \left(\frac{K'_i}{K'_o}\right)^{1/b}\right]^b} \sum_{q=1}^{n} \delta^{b-1}{}_q \cos^2\left(\frac{2\pi q}{n}\right) \tag{200}$$

If b is unity

$$F_{yu} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \delta_q \sin\frac{2\pi q}{n} \tag{201}$$

$$F_{zu} = \frac{K'_i}{\left[1 + \frac{K'_i}{K'_o}\right]} \sum_{q=1}^{n} \delta_q \cos\frac{2\pi q}{n} \tag{202}$$

The associated derivatives are given by Equations (188) through (191).

Special Cases of Initial Constraint. Up to this point it was assumed that all outer rings were rigidly attached to a flexible supporting structure. There are many instances where one or more outer rings are free from constraint in one or more degrees of freedom or where the constraints are different than those dictated by the structural flexibilities.

Some of the more common cases are treated in the following.

Single Ball Bearings With Axially Floating OD's. In a system having u bearings let the vth ball bearing be so mounted that its OD is free to slide axially without restraint.

The matrix of influence coefficients $[{}_1\Delta_{z_u}]$ should contain zeros in the vth row and column.

Let there have been made an assumption of the five displacements x_B, y_B, z_B, ϑ_{y_B}, and ϑ_{z_B} from which have been obtained $\bar{y}_{u=v}$, $\bar{z}_{u=v}$, $\bar{\alpha}_{yu=v}$, and $\bar{\alpha}_{zu=v}$. Then the value of $\bar{x}_{u=v}$ is that which satisfies the relation

$$F_{xu=v} = 0 = \Phi \tag{203}$$

If $\bar{x}^\circ_{u=v}$ is an initial estimate of $\bar{x}_{u=v}$, an improved estimate is $\bar{x}'_{u=v}$, where

$$\bar{x}_{u=v}{}' = \bar{x}_{u=v}{}^\circ - \frac{\Phi}{\left(\frac{dF_{xu=v}}{d\bar{x}_{u=v}}\right)} \tag{204}$$

The derivative required in Equation (204) is given by Equa-

tion (91) with $\delta_u = \bar{x}_{u=v}$ for the dynamic case or by Equation (135) for the static case.

The converged value of $\bar{x}_{u=v}$, together with $\bar{y}_{u=v}$, $\bar{z}_{u=v}$, $\bar{\alpha}_{y_{u=v}}$, and $\bar{\alpha}_{z_{u=v}}$ are used to obtain new estimates of x_B, y_B, z_B, ϑ_{y_B}, and ϑ_{z_B}. During the latter process only all derivatives with respect to $\bar{x}_{u=v}$ are set equal to zero.

Single Ball Bearings With a Spherical Outer Ring OD. If the vth ball bearing has its outer ring mounted in a spherical seat the moment reaction at that bearing is zero.

The influence-coefficient matrices $[{}_1\alpha_{ru}]$, $[{}_1\alpha'_{ru}]$, $[{}_1\alpha_{su}]$, and $[{}_1\alpha'_{su}]$ should contain zeroes in the vth row and column.

For an assumption of the five displacements at point B values of $\bar{x}_{u=v}$, $\bar{y}_{u=v}$, and $\bar{z}_{u=v}$ are obtained directly.

The values of $\bar{\alpha}_{y_{u=v}}$ and $\bar{\alpha}_{z_{u=v}}$ are those which satisfy the relations

$$M_{y_{u=v}} = 0 = \chi_1 \tag{205}$$

$$M_{z_{u=v}} = 0 = \chi_2 \tag{206}$$

Let $\bar{\alpha}^\circ_j$ be initial estimates of $\bar{\alpha}_{y_{u=v}}$ and $\bar{\alpha}_{z_{u=v}}$ where $j = 1, 2$, respectively. Improved estimates are $\bar{\alpha}_j'$, where:

$$\bar{\alpha}_j' = \bar{\alpha}_j^\circ - [C_{ij}]^{-1}\{\chi_i\} \tag{207}$$

where:

$$c_{11} = \frac{\partial M_{y_{u=v}}}{\partial \bar{\alpha}_{y_{u=v}}} \tag{208}$$

$$c_{12} = \frac{\partial M_{y_{u=v}}}{\partial \bar{\alpha}_{z_{u=v}}} \tag{209}$$

$$c_{21} = \frac{\partial M_{z_{u=v}}}{\partial \bar{\alpha}_{y_{u=v}}} \tag{210}$$

$$c_{22} = \frac{\partial M_{z_{u=v}}}{\partial \bar{\alpha}_{z_{u=v}}} \tag{211}$$

The converged values of $\bar{\alpha}_{y_{u=v}}$ and $\bar{\alpha}_{z_{u=v}}$ are used with $\bar{x}_{u=v}$, $\bar{y}_{u=v}$, and $\bar{z}_{u=v}$ to obtain new estimates of the displacements at point B, and so forth. In the latter process only all derivatives with respect to $\bar{\alpha}_{y_{u=v}}$ and $\bar{\alpha}_{z_{u=v}}$ are set equal to zero.

Multiplex Bearings Which Act as a Unit and Whose ODs Are Free to Float Axially. Consider a system of bearings in which v ball bearings are so arranged that their outer races are axially fixed relative to each other but free to move axially in their housings without restraint.

Let the relative axial motion of any inner race with respect to its outer race be x. Then the total axial displacement of an inner race with respect to its outer race is $\bar{x}_{u=v}$.

$$\bar{x}_{u=v} = x + \Delta''_{x_{u=v}} \tag{212}$$

Assume that an initial choice has been made of the displacements at point B from which have been calculated $\bar{y}_{u=v}$, $\bar{z}_{u=v}$, $\bar{\alpha}_{y_{u=v}}$, and $\bar{\alpha}_{z_{u=v}}$. Then the value of x is that which satisfies the relation

$$\sum_v F_{x_{u=v}} = 0 = \xi \tag{213}$$

For an initial estimate, x°, of x an improved estimate is x' where

$$x' = x^\circ - \frac{\xi}{\sum_v \left(\frac{\partial F_{x_{u=v}}}{\partial \bar{x}_{u=v}}\right)} \tag{214}$$

The converged value of x enables new estimates of the displacements at B, and so forth.

Calculation of Ball Bearing Fatigue Life. The Lundberg-Palmgren theory, Reference [2], is the basis for the AFBMA standard method for evaluating load ratings of ball bearings and is used by most bearing manufacturers in establishing catalog ratings. However, this theory is based upon the assumption that parallel ring displacement exists, that is, that the displacements $\bar{\alpha}_{y_u}$ and $\bar{\alpha}_{z_u}$ are zero. Also, it assumes that, under combined load, the operating contact angle is the same at each ball position.

The foregoing solution of the elastic problem yields directly the true ball load-distribution and contact-angle variation. By evaluating the fatigue effects of the ball passages over an exact path, a more accurate estimate of fatigue life is possible.

Equation (115) of Reference [2] gives the capacity of a race contact for 10^6 revolutions of the inner race and 90 per cent probability of survival.

Recent work, as yet incomplete, has shown that the spin of the ball relative to a raceway has an important effect on fatigue life. Although this effect has not been fully evaluated, preliminary results indicate that an exponential relation exists between the capacity of a race contact and the ratio $\frac{\omega_{sq}}{\omega_{R_q}}$. Therefore, Equation (115) of Reference [2] should be modified as follows

$$Q_{c_q} = A\left[1 - k\left|\frac{\omega_{s_q}}{\omega_{R_q}}\right|\right]^{\nu}\left[\frac{2f}{2f-1}\right]^{.41} \frac{[1 \mp \gamma\cos\beta_q]^{1.39}}{[1 \pm \gamma\cos\beta_q]^{1/3}}\,\gamma^{.3}d^{.8}n^{-1/3} \tag{215}$$

The upper sign is used for an inner-race contact and the lower sign for an outer-race contact.

For a raceway in which a particular point supports the same load at each ball passage the life of the race in hours for 90 per cent probability of survival is

$${}_BL_{1u} = \frac{10^6}{60N\left[\frac{1}{n}\sum_{q=1}^{n}\left(\frac{P_q}{Q_{c_q}}\right)^{10/3}\right]^{3}} \tag{216}$$

For a raceway in which the loading of a particular point varies cyclically with time, the life of the raceway in hours for 90 per cent probability of survival is

$${}_BL_{2u} = \frac{10^6 n}{60N\sum_{q=1}^{n}\left(\frac{P_q}{Q_{c_q}}\right)^{3}} \tag{217}$$

The life of a complete ball bearing is

$${}_BL_u = \left[\left(\frac{1}{{}_BL_{1u}}\right)^{10/9} + \left(\frac{1}{{}_BL_{2u}}\right)^{10/9}\right]^{-.9} \tag{218}$$

The ratio $\frac{\omega_{sq}}{\omega_{R_q}}$, required in Equation (215) is, for an outer-race contact

$$\frac{\omega_{s_{o_q}}}{\omega_{R_{o_q}}} = -(1 + \gamma\cos\beta_{o_q})\tan(\beta_{o_q} - \alpha_q) + \gamma\sin\beta_{o_q} \tag{219}$$

and, for an inner-race contact

$$\frac{\omega_{s_{i_q}}}{\omega_{R_{i_q}}} = (1 - \gamma\cos\beta_{i_q})\tan(\beta_{i_q} - \alpha_q) + \gamma\sin\beta_{i_q} \tag{220}$$

Calculation of Roller Bearing Fatigue Life. The calculation of the fatigue life of a radial roller bearing is greatly dependent upon the manner in which the rollers are crowned. Under light loads a continuously crowned roller may have substantially point con-

tact and the pressure area is elliptical. Under increasing load the pressure area encompasses the whole length of the roller and, at a certain value of load, the pressure distribution is substantially uniform along the roller's effective length. At higher loading, concentrations of stress occur at the roll ends. For substantially light loads corresponding to elliptical contact areas (point contact) life is inversely proportional to the third power of the load. For heavier loads corresponding to line contact, life is inversely proportional to the fourth power of the load. Unequal load distribution along the roller length introduces further complications.

Most aircraft roller bearings usually have rollers which are cylindrical for about $^2/_3$ of their length. Crowning is accomplished by tapering the roll ends slightly for a distance approximately $^1/_3$ of the roller length from each end. Such bearings are relatively free from the harmful effects of edge loading at all loads, provided alignment is good. Since substantially line contact exists, the life is inversely proportional to the fourth power of the load. The discussion of roller-bearing life is restricted to bearings of this type.

Equation (116a) of Reference [3] gives the capacity of a race contact for 90 per cent probability of survival to 10^6 revolutions of the inner race, for line contact, as

$$Q_c' = A' \frac{[1 \mp \gamma]^{29/27}}{[1 \pm \gamma]^{1/4}} \gamma^{3/4} d^{29/27} l_e^{7/8} n^{-1/4} \qquad (221)$$

The upper signs are used with an inner raceway and the lower with the outer raceway.

For a raceway in which a particular point supports the same load at each roll passage the life of the race in hours for 90 per cent probability of survival is

$${}_RL'_{1a} = \frac{10^6}{60N\left[\frac{1}{n}\sum_{q=1}^{n}\left(\frac{P_q}{Q_c'}\right)^{9/2}\right]^{2/9}} \qquad (222)$$

For a raceway in which the loading of a particular point varies cyclically with time, the life of the raceway in hours for 90 per cent probability of survival is

$${}_RL'_{2a} = \frac{10^6 n}{60N\sum_{q=1}^{n}\left(\frac{P_q}{Q_c}\right)^4} \qquad (223)$$

The life of the complete roller bearing is

$${}_RL'_a = \left[\left(\frac{1}{{}_RL'_{1a}}\right)^{9/8} + \left(\frac{1}{{}_RL'_{2a}}\right)^{9/8}\right]^{-8/9} \qquad (224)$$

References

1 A. B. Jones, "Analysis of Stresses and Deflections," vols. 1 and 2, New Departure Division, GMC, Bristol, Conn., 1946.

2 G. Lundberg and A. Palmgren, "Dynamic Capacity of Rolling Bearings," *Acta Polytechnica, Mechanical Engineering Series*, vol. 1, no. 3, Stockholm, Sweden, 1947.

3 G. Lundberg and A. Palmgren, "Dynamic Capacity of Roller Bearings," *Acta Polytechnica, Mechanical Engineering Series*, vol. 2, no. 4, Stockholm, Sweden, 1952.

DISCUSSION

T. Tallian[2]

The paper deals with a subject of considerable practical interest. No general method for obtaining analytically the forces and elastic deflections in a redundant system of radial ball and roller bearings, supported by elastic members has been published before, as far as I am able to ascertain. Yet the problem, and several special cases thereof, arises frequently and the analyst seeking a solution to it will highly welcome the aid afforded by a generally applicable method. By making the IBM computer program available to the public, the author could further ensure general usefulness of the work done.

With respect to the fatigue life calculations outlined in Sections entitled, "Calculation of Ball Bearing Fatigue Life" and "Calculation of Roller Bearing Fatigue Life," p. 319, a word of caution seems appropriate. The Lundberg-Palmgren theory, given as Reference [2] of the paper, is based on the assumption of pure normal loading of all rolling contacts in the absence of a hydrodynamic lubricant film affecting the stress distribution in the contact. Its applicability to practice, as in the AFBMA standard load rating method, is ensured by the empirical fact that, for the usual operating conditions, the actual (but analytically inaccessible) stress conditions are proportional in their fatigue effects to the theoretically used stress conditions. Empirical factors of proportionality give correct orders of life length.

Attempts to improve the precision of the Lundberg-Palmgren theory or to extend it to operating conditions for which it was not empirically verified, should, in my opinion, be directed toward better definition of the stress conditions actually existing. If, instead, one applies correction factors to the end-results as in Equation (215), then the formulas obtained become purely empirical, and are no longer a logical outgrowth of the theory. They are interpolation formulas only, applicable to conditions experimentally elucidated.

As an example of cases where experimental results contradict an extrapolated theory, I point to some experiments indicating that, under certain conditions of slip at the contacts, the mechanism of failure changes from fatigue flaking to surface damage. Clearly, a fatigue theory cannot carry over to operating conditions where such occurs.

Regarding the Section on Roller Bearing Fatigue Life, similar remarks may be made. It is, e.g., experimentally established that crowned roller bearings will not behave as predicted by a Hertzian fatigue theory if lubrication is marginal. It appears that, in that case, lubrication breaks down at the roller ends even though the roller is crowned and this results in unexpected failures.

[2] SKF Industries, Inc., Philadelphia, Pa.

A Sommerfeld Solution for Finite Bearings With Circumferential Grooves[1]

J. V. FEDOR
Naval Ordnance Laboratory,
Silver Spring, Md.
Assoc. Mem. ASME

A method of solution is developed that circumvents the algebraic complexity in the solution of Reynolds differential equation applied to full journal bearings. The method leads to equations for journal-bearing characteristics that are in finite form. Salient features of the complete-oil-film solution are: As the eccentricity ratio b approaches 1, the load capacities of all finite bearings approach that of the infinite bearing; also, the friction curve intercept, $(r/c)f$, is 1 for all finite bearings when b equals 1. Results are compared with calculated values published by Muskat and Morgan for $0 < b < 0.6$. Excellent agreement is found throughout the compared range.

IT HAS BEEN a long-standing belief of many scientists and engineers in the lubrication field that it is unlikely that a computable analytical solution to Reynolds three-dimensional equation for the full journal bearing would be obtained. If papers by Muskat and Morgan [1][2] and Jacobson [2] are examined, one is impressed by the complex nature of even approximate solutions to Reynolds equation. When one applies classical methods of solution (separation of variables), the justification for this belief becomes apparent. A formal classical solution for Reynolds equation is constructed in Ref. [9] and it will not be repeated here.

The question of boundary conditions has compounded the problem further (see [3] and [4]). It is generally agreed that, with an adequate oil supply and relatively light loads, it is possible to provide sufficient oil to form a complete oil film. This forms the basis for the Sommerfeld boundary condition; that is, the pressure and its first derivative must duplicate themselves when going from 0 to 2π. The Sommerfeld boundary condition will be used in this paper because of the mathematical simplicity that it produces. It is the purpose of the following section to outline the journal-bearing problem and present a method of solution which circumvents the algebraic complexities that arise in the solution of Reynolds equation.

Analytical Solution of Reynolds Equation

Reynolds equation for the pressure distribution in a full journal bearing is

$$\frac{\partial}{\partial\theta}\left[\frac{(1+b\cos\theta)^3}{\mu}\frac{\partial P}{\partial\theta}\right]+r^2\frac{\partial}{\partial z}\left[\frac{(1+b\cos\theta)^3}{\mu}\frac{\partial P}{\partial z}\right]=-\frac{6Urb}{c^2}\sin\theta \quad (1)$$

where

P = pressure at any point in oil film above atmospheric pressure
U = linear velocity of journal
r = radius of journal
c = difference between journal and bearing radii
e = eccentricity of journal and bearing axes
μ = viscosity of oil-constant
h = oil film thickness $= c(1 + b\cos\theta)$
b = eccentricity ratio $= e/c$
θ = angular co-ordinate
z = co-ordinate in axial direction

The co-ordinate system shown in Fig. 1 will be used in this analysis.

Putting a circumferential groove into a bearing is equivalent to cutting the bearing in two parts and providing each part with a uniform lubricant pressure around the adjacent ends. The pressure distribution in the interior of the two parts will be independent of each other and each will behave as an independent bearing of length determined by the total length of the bearing and the position of the groove. Therefore it is necessary to consider only end-fed journal bearings when dealing with bearings with circum-

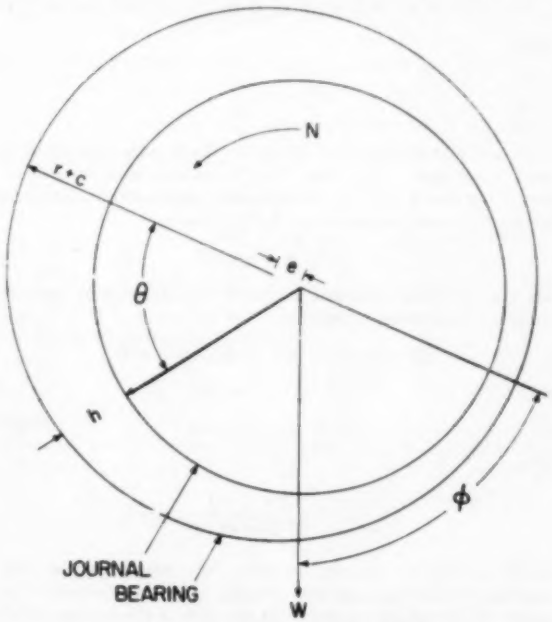


Fig. 1 Cross section of journal bearing

[1] Part of this paper was prepared by the author in partial fulfillment of requirements for the degree of Doctor of Philosophy, Mechanical Engineering, University of Maryland, College Park, Md.

[2] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

NOTE: Statements and opinions advanced in papers are to be considered as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, December 29, 1958. Paper No. 59—Lub-4.

ferential grooves. For convenience, a journal bearing with a groove located at the center of the bearing will be discussed, see Fig. 2. In the axial direction, the origin of the co-ordinate system will be located midway in the end-fed bearing. It is assumed that the width of the groove is small compared to l.

First the pressure distribution must be found in the bearing by solving Reynolds equation. The journal-bearing characteristics are then obtained by integration. The following boundary conditions will be imposed on the pressure equation:

1 The pressure function and its first derivative must duplicate themselves in going completely around the bearing; i.e.,

$$P(0, z) = P(2\pi, z) \quad \text{and} \quad \frac{\partial P(0, z)}{\partial \theta} = \frac{\partial P(2\pi, z)}{\partial \theta}$$

2 The pressure function must vanish at the end of the bearing: $P(\theta, l/2) = 0$; and at $z = -l/2$, the pressure must be P_0, the pressure of the circumferential groove.

For the total pressure function let

$$P(\theta, z) = P_0\left(\frac{1}{2} - z/l\right) + \frac{6\mu U r b[2 + b\cos\theta]\sin\theta}{c^2(2 + b^2)[1 + b\cos\theta]^2} - \sum_1^\infty A_m(z)\sin m\theta \quad (2)$$

It will be noticed that Equation (2) meets the first boundary condition. The first term represents the circumferential source function, the second term is the pressure equation for the infinite bearing, $P_\infty(\theta)$, and the third term represents a correction function to take into account the finiteness of the bearing. It will be readily verified that the first two terms satisfy Reynolds equation. After equating coefficients of $\sin m\theta$ to zero in Reynolds equation, the following recurrence formulas are obtained from the third term

$$2(D^2 - 1)A_1(z) + b(D^2 + 2)A_2(z) = 0 \quad (3)$$

and

$$2(D^2 - m^2)A_m(z) + b(D^2 - m^2 - m + 2)A_{m-1}(z) + b(D^2 - m^2 + m + 2)A_{m+1}(z) = 0 \quad m > 1 \quad (4)$$

where

$$D^2 = r^2\frac{d^2}{dz^2}$$

At first glance Equations (3) and (4) look quite formidable to solve. Not only is Equation (4) a three-term recurrence equation, but it also is a differential equation. As a gambit suppose we let $A_2(z)$ be some proportion of $A_1(z)$. That is,

$$A_2(z) = -\epsilon A_1(z) \quad (5)$$

where ϵ is a positive quantity to be determined. Making this substitution, Equation (3) becomes

$$[(2 - \epsilon b)D^2 - 2(1 + \epsilon b)]A_1(z) = 0$$

or

$$(D^2 - \alpha^2)A_1(z) = 0 \quad (6)$$

where

$$\alpha^2 = \frac{2(1 + \epsilon b)}{2 - \epsilon b}$$

Equation (6) is a simple second-order differential equation with constant coefficients and the solution is straightforward. To satisfy the boundary conditions at the ends of the bearing, $A_1(z)$ must be even in z. Hence we take

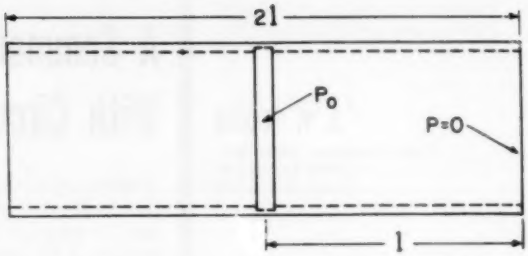


Fig. 2 Diagrammatic representation of bearing with circumferential groove

$$A_1(z) = C_1\cosh\alpha z/r \quad (7)$$

where C_1 is an arbitrary constant.

It follows from the initial assumption that $A_2(z)$ is given by

$$A_2(z) = -\epsilon C_1\cosh\alpha z/r \quad (8)$$

$A_3(z)$ is determined from the equation

$$2(D^2 - 4)A_2(z) + b(D^2 - 4)A_1(z) + bD^2A_3(z) = 0 \quad (9)$$

and $A_4(z)$ from

$$2(D^2 - 9)A_3(z) + b(D^2 - 10)A_2(z) + b(D^2 - 4)A_4(z) = 0 \quad (10)$$

Solving for $A_3(z)$ and $A_4(z)$ yields

$$A_3(z) = \left(\frac{2\epsilon}{b} - 1\right)\frac{(\alpha^2 - 4)}{\alpha^2}C_1\cosh\alpha z/r + C_3 \quad (11)$$

and

$$A_4(z) = -\frac{2}{b}\frac{(\alpha^2 - 9)}{\alpha^2}\left(\frac{2\epsilon}{b} - 1\right)C_1\cosh\alpha z/r - \frac{9C_3}{2b} + \frac{\epsilon(\alpha^2 - 10)C_1}{\alpha^2 - 4}\cosh\alpha z/r + C_4\cosh 2z/r \quad (12)$$

In general $A_{m+1}(z)$ can be evaluated from the recurrence formula

$$(D^2 - m^2 + m + 2)A_{m+1}(z) = -\frac{2}{b}(D^2 - m^2)A_m(z) - (D^2 - m^2 - m + 2)A_{m-1}(z) \quad (13)$$

Since Equation (13) is a second-order differential equation for $A_{m+1}(z)$, there will always be constants available in each $A_{m+1}(z)$ to satisfy the boundary conditions at the ends of the bearing.

The remaining boundary conditions will now be applied: At the end of the bearing ($z = l/2$) the pressure function must vanish

$$0 = \frac{6\mu U r b[2 + b\cos\theta]\sin\theta}{c^2(2 + b^2)[1 + b\cos\theta]^2} - \sum_1^\infty A_m(l/2)\sin m\theta \quad (14)$$

This requires that the first term, $P_\infty(\theta)$, in Equation (14) be expanded in a Fourier series. This can be done by noting the $P_\infty(\theta)$ can be separated into two terms and each term expanded into a Fourier series

$$P_\infty(\theta) = \frac{6\mu U r}{c^2(2 + b^2)}\left\{\frac{b}{(1 + b\cos\theta)} + \frac{b}{(1 + b\cos\theta)^2}\right\}\sin\theta \quad (15)$$

It is a well-known fact that $\dfrac{\sin\theta}{1 + 2a\cos\theta + a^2}$ can be expanded in a Fourier series [5]

$$\frac{\sin\theta}{1+2a\cos\theta+a^2}=\sum_1^\infty(-a)^{m-1}\sin m\theta$$

which is uniformly convergent for $0<a<1$. If the first term of Equation (15) is taken and let

$$1/b=\frac{a+a^{-1}}{2}$$

then

$$\frac{b\sin\theta}{1+b\cos\theta}=\frac{2a\sin\theta}{1+2a\cos\theta+a^2}=2\sum_1^\infty(-)^{m-1}a^m\sin m\theta \quad (16)$$

Differentiating Equation (16) with respect to b results in

$$\frac{\partial}{\partial b}\left(\frac{b\sin\theta}{1+b\cos\theta}\right)=\frac{b\sin\theta}{(1+b\cos\theta)^2}=2m\sum_1^\infty\frac{(-)^{m-1}a^m\sin m\theta}{(1-b^2)^{1/2}} \quad (17)$$

Combining Equations (16) and (17) the Fourier series expansion of $P_\infty(\theta)$ is obtained

$$P_\infty(\theta)=\frac{12\mu Ur}{c^2(2+b^2)}\sum_1^\infty(-)^{m-1}a^m\left[1+\frac{m}{(1-b^2)^{1/2}}\right]\sin m\theta \quad (18)$$

where

$$a=\frac{b}{1+(1-b^2)^{1/2}}$$

Substituting this into Equation (14) leads to

$$0=\frac{12\mu Ur}{c^2(2+b^2)}\sum_1^\infty(-)^{m-1}a^m\left[1+\frac{m}{(1-b^2)^{1/2}}\right]\sin m\theta-\sum_1^\infty A_m(l/2)\sin m\theta \quad (19)$$

Equating coefficients of $\sin m\theta$ results in the following

$$A_m(l/2)=\frac{12\mu Ur(-)^{m-1}a^m}{c^2(2+b^2)}\left[1+\frac{m}{(1-b^2)^{1/2}}\right] \quad (20)$$

Now ϵ [from the assumption $A_2(z)=-\epsilon A_1(z)$] can be evaluated by taking the first two terms of Equation (20)

$$A_1(l/2)=C_1\cosh\alpha\beta=\frac{12\mu Ura}{c^2(2+b^2)}\left[1+\frac{1}{(1-b^2)^{1/2}}\right] \quad (21)$$

$$A_2(l/2)=-\epsilon C_1\cosh\alpha\beta=-\frac{12\mu Ura^2}{c^2(2+b^2)}\left[1+\frac{2}{(1-b^2)^{1/2}}\right] \quad (22)$$

where $\beta=l/2r$

Dividing one equation by the other gives

$$\epsilon=\frac{a[2+(1-b^2)^{1/2}]}{[1+(1-b^2)^{1/2}]}=\frac{b[2+(1-b^2)^{1/2}]}{[1+(1-b^2)^{1/2}]^2} \quad (23)$$

It is thus seen that ϵ is evaluated by the boundary condition at the end of the bearing. It is readily verified that the pressure function [Equation (2)] reduces to P_0 when $z=-l/2$.

In the following section the convergence of the series in Equation (2) is demonstrated. Since Equation (2) satisfies Reynolds equation and meets the boundary conditions specified, it is a solution of the problem under consideration. It will be noticed that the method used to solve Reynolds equation does not require the use of any expansion theorem except that connected with Fourier series. Thus the method is simple and easy to apply. In addition, it will be seen that this method leads to simple equations for the journal-bearing characteristics.

Convergence of Series

It will now be shown that the series

$$\sum_1^\infty A_m(z)\sin m\theta$$

converges in the region under consideration. This will be done by getting an approximate expression for $A_m(z)$ for large m and investigating convergence. Making the approximation

$$(m-1)(m+2)\simeq(m-1)^2$$

$$(m+1)(m-2)\simeq(m+1)^2$$

for large m, Equation (4) becomes

$$2[D^2-m^2]A_m(z)+b[D^2-(m-1)^2]A_{m-1}(z)+b[D^2-(m+1)^2]A_{m+1}(z)=0 \quad (24)$$

Let

$$[D^2-m^2]A_m(z)=F_m \quad (25)$$

where F_m is a function of m, b, and z. Substituting into Equation (24) gives the following second-order difference equation with constant coefficients

$$2F_m+bF_{m-1}+bF_{m+1}=0 \quad (26)$$

It will be verified that a solution of Equation (26) finite at $b=0$ is

$$F_m=(-)^m\left(\frac{b}{1+(1-b^2)^{1/2}}\right)^m f(z)=(-)^m a^m f(z) \quad (27)$$

where $f(z)$ is an even function of z independent of m. Equation (25) now becomes

$$[D^2-m^2]A_m(z)=(-)^m a^m f(z) \quad (28)$$

An even solution of Equation (28) is

$$A_m(z)=C_m\cosh mz/r+\frac{(-)^m a^m}{mr}\int f(z')\sinh\frac{m}{r}(z-z')dz' \quad (29)$$

It is reasonable to assume that $f(z)$ is bounded for $0<b<1$ and z finite. If $f(z)$ is replaced by its maximum value M, the right-hand side of Equation (29) becomes

$$C_m\cosh mz/r-\frac{(-)^m a^m M}{m^2} \quad (30)$$

The constant C_m can be evaluated from the boundary condition at the end of the bearing. At $z=l/2$,

$$A_m(l/2)=C_m\cosh m\beta-\frac{(-)^m a^m M}{m^2}$$

$$C_m=\frac{A_m(l/2)+\dfrac{(-)^m a^m M}{m^2}}{\cosh m\beta} \quad (31)$$

Equation (30) becomes

$$\frac{A_m(l/2)\cosh mz/r}{\cosh m\beta} + \frac{(-)^m a^m M}{m^2}\left(\frac{\cosh mz/r}{\cosh m\beta} - 1\right) \quad (32)$$

Taking the absolute value of the terms in Equation (32) the following inequality exists for large m:

$$|A_m(z)| \leq \left|\frac{A_m(l/2)\cosh mz}{\cosh m\beta}\right| + \frac{a^m M}{m^2}\left|\frac{\cosh mz/r}{\cosh m\beta} - 1\right| \quad (33)$$

or

$$|A_m(z)| \leq |A_m(l/2)| + \left|\frac{a^m M}{m^2}\right| \quad (34)$$

where the inequality has been improved by neglecting some of the terms.

Denoting the right-hand side of Equation (34) by B_m, it is noted by the Cauchy ratio test that the series formed by B_m

$$\sum_1^\infty B_m$$

is absolutely convergent since a is less than 1 for $0 < b < 1$.

Since $|\sin m\theta| \leq 1$ and $|A_m(z)| \leq B_m$ for large m, it follows from the comparison test that the series

$$\sum_1^\infty A_m(z)\sin m\theta$$

absolutely converges for $-l/2 \leq z \leq l/2$, $0 \leq \theta \leq 2\pi$, and $0 < b < 1$.

Journal-Bearing Characteristics

Load Capacity. All of the $A_m(z)$ can be calculated from the recurrence formula Equation (13), but this is not necessary when determining the characteristics of journal bearings operating with the Sommerfeld boundary condition. The load components are given by

$$W\sin\varphi = \int_{-l/2}^{l/2}\int_0^{2\pi} P(\theta, z)\sin\theta\, rd\theta dz \quad (35a)$$

and

$$W\cos\varphi = -\int_{-l/2}^{l/2}\int_0^{2\pi} P(\theta, z)\cos\theta\, rd\theta dz \quad (35b)$$

By direct integration, noting the orthogonality of $\sin m\theta$ and $\cos m\theta$, it will be found that

$$W\cos\varphi = 0$$

that is, $\varphi = \pi/2$ and W has the very simple form

$$W = W_\infty\left(1 - \frac{\tanh\alpha\beta}{\alpha\beta}\right) = W_\infty K \quad (36)$$

where

$$W_\infty = \frac{12\mu U r^2 bl\pi}{c^2(2 + b^2)(1 - b^2)^{1/2}}, \quad K = 1 - \frac{\tanh\alpha\beta}{\alpha\beta} \quad (37)$$

and

$$\alpha^2 = \frac{2(1 + \epsilon b)}{2 - \epsilon b}, \quad \epsilon = \frac{b[2 + (1 - b^2)^{1/2}]}{[1 + (1 - b^2)^{1/2}]^2}$$

Notice that as $\beta \to \infty$, the load capacity goes over to the Sommerfeld load capacity for the infinite bearing as it should.

Coefficient of Friction. The shear stress acting on the journal-bearing surface [6] is given by

$$\sigma = \frac{\mu U}{h} \pm \frac{h}{2r}\frac{\partial P}{\partial\theta} \quad (38)$$

where the + refers to the traction on the journal and the − to that on the bearing. Substituting $h = c(1 + b\cos\theta)$ and integrating the right-hand side over the journal or bearing surface gives the friction force

$$F = \int_{-l/2}^{l/2}\int_0^{2\pi}\left\{\frac{\mu U}{c(1 + b\cos\theta)} \pm \frac{c(1 + b\cos\theta)}{2r}\frac{\partial P}{\partial\theta}\right\} rd\theta dz \quad (39)$$

Integrating the first term completely and integrating the second term by parts and noting the boundary condition $P(0, z) = P(2\pi, z)$ results in

$$F = \frac{2\pi\mu U r l}{c(1 - b^2)^{1/2}} \pm \frac{cb}{2r}\int_{-l/2}^{l/2}\int_0^{2\pi} P(\theta, z)\sin\theta\, rd\theta dz \quad (40)$$

Now the integral is simply the component of load perpendicular to the line of centers, $W\sin\varphi$. The coefficient of friction is defined as the ratio of the friction force to the load capacity of the bearing. Dividing by the load W, rearranging, noting that $\varphi = \pi/2$, and using Equation (36) the coefficient of friction is given by

$$\left(\frac{r}{c}\right) f = \frac{2 + b^2}{6bK} \pm b/2 \quad (41)$$

A convenient dimensionless quantity often used in journal-bearing analysis is the Sommerfeld variable S. This is defined as

$$S = \left(\frac{r}{c}\right)^2 \frac{\mu N}{W'} \quad (42)$$

where W' is the load per unit of projected area and N is the angular speed of the journal. Substituting Equation (36) into (42) results in

$$S = \frac{(2 + b^2)(1 - b^2)^{1/2}}{12\pi^2 bK} \quad (43)$$

End Oil Flow. Reference [6] gives the following expressions for the flow in the circumferential and axial directions for unit width in a direction perpendicular to the flow

$$q_\theta = \frac{Uh}{2} - \frac{h^3}{12\mu r}\frac{\partial P}{\partial\theta} \quad (44a)$$

$$q_z = -\frac{h^3}{12\mu}\frac{\partial P}{\partial z} \quad (44b)$$

The axial flow (end leakage) is obtained by integrating Equation (44b) over the circumference of the bearing at the end of the bearing. Hence the oil flow out of one end of the bearing is

$$Q = -\int_0^{2\pi}\frac{h^3}{12\mu}\frac{\partial P}{\partial z}\Big]_{-l/2} rd\theta \quad (45)$$

Substituting Equation (2) into (45) and carrying out the integration leads to

$$Q_s = \frac{\pi P_0 c^3(1 + 3b^2/2)}{12\mu\beta} \quad (46)$$

or in dimensionless form

$$\frac{Q_s\mu}{P_0 c^3} = \frac{\pi(1 + 3b^2/2)}{12\beta} \quad (47)$$

The subscript is used to indicate that the flow is due entirely to the circumferential groove.

To insure that the oil film is continuous, the condition is imposed that the axial pressure gradient at the end of the bearing be negative [1]

$$\left.\frac{\partial P}{\partial z}\right]_{z=l/2} = -\frac{P_0}{l} - \sum_1^\infty \frac{\partial A_m}{\partial z} \sin m\theta_m < 0 \qquad (48)$$

where θ_m is the angle at which the tendency for positive gradients is the greatest. Equation (48) insures that the velocity of the oil is positive (outward) at the end of the bearing. Taking the first term of the series and rearranging, a simple equation is arrived at that imposes a minimum value for the feed-pressure parameter

$$\frac{\left(\frac{c}{r}\right)^2 P_0}{\mu N} \geq \frac{48\pi b\alpha\beta \tanh \alpha\beta}{(2+b^2)(1-b^2)^{1/2}} \qquad (49)$$

It is clear from Equation (49) that for a given eccentricity and length-to-diameter ratio that a continuous oil film depends not only on the feed pressure but also on the viscosity, speed of the journal, and the ratio of clearance to radius of the journal (all of these quantities are related to the load capacity). Note that the left-hand side of Equation (49) is similar to the reciprocal of the Sommerfeld variable with the feed pressure P_0 replacing the load per projected area W'.

Temperature and Viscosity of the Oil. A simple method suggested by Wilcock and Rosenblatt [7] for estimating the outlet temperature and viscosity of the oil of a bearing is to equate the power dissipated in keeping the journal rotating to the energy (heat) picked up by the oil flowing out of the bearing. This gives an expression for the viscosity in terms of the outlet temperature T_2 and other parameters of the journal-bearing system. A plot of μ versus T_2 is then constructed. Superposed on this is a plot of μ versus T of the oil used. The intersection of these curves gives the value of the outlet viscosity and temperature. If the journal bearing is considered as a "mixer" where some oil at viscosity μ_1 goes in and some oil with a viscosity μ_2 goes out, then μ_2 can be considered as the viscosity of the oil in the bearing. This viscosity μ_2 is used to calculate the power loss of the bearing and the load of the bearing.

The power dissipated (H) due to friction is equal to the product of the friction force on the journal times the peripheral speed of the journal

$$H = FU = fWU \qquad (50)$$

The energy acquired by the oil flowing out of the bearing is

$$Q\gamma C_p \Delta T$$

where γ is the weight density of the oil, C_p is the specific heat, and ΔT is the temperature rise. Equating the two, rearranging, and taking into account the difference in units (J), an expression is obtained for μ in terms of the parameters of the bearing

$$\mu^2 = \frac{JP_0\left(\frac{c}{r}\right)^2\left(1+\frac{3b^2}{2}\right)\gamma C_p \Delta T S}{96\beta^2 N^2 \left(\frac{r}{c}\right) f} \qquad (51)$$

Values for S and $\left(\frac{r}{c}\right)f$ are obtained from Equations (41) and (42) for a fixed value of b.

The bearing characteristics have been determined. It is seen that by modifying the classical method of separation of variables, a solution is obtained for Reynolds equation which leads to simple equations for the characteristics of a journal bearing with circumferential groove and complete oil film.

Discussion of Results and Conclusions

Discussion. Now that an exact solution for the full journal bearing with a continuous oil film has been obtained, it is interesting to examine the results in regard to friction and load capacity. In addition, compare results with other published solutions.

Fig. 3 shows a plot of $(r/c)f$ versus S for a few length-to-diameter ratios. A Petroff line [$(r/c)f = 2\pi^2 S$] is also included for comparison. A salient point of the curves is the ordinate intercept. All of the complete-oil-film friction curves for journal intercept the $(r/c)f$ axis at 1. Examination of Equation (41) shows that all friction curves for the bearing go through the origin when b is equal to 1 ($S = 0$).

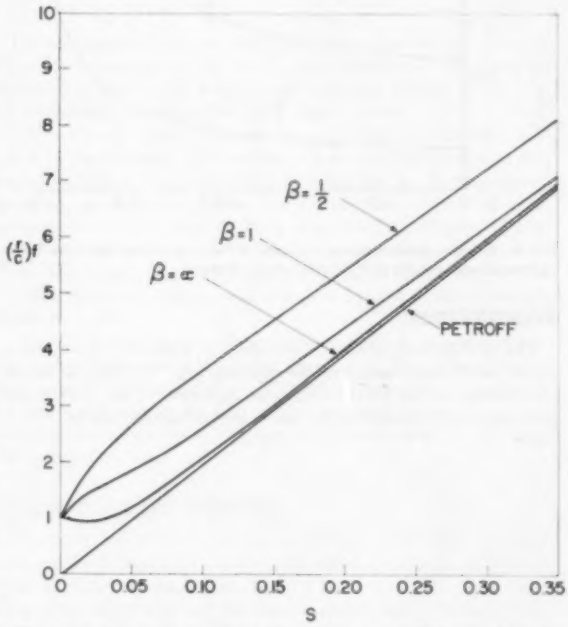


Fig. 3 Coefficient of friction $(r/c)f$ versus Sommerfeld variable S for journal

Fig. 4 compares leakage factors (W/W_∞) obtained from the exact continuous-oil-film solution with results published by Muskat and Morgan [1] for values of b up to 0.6. Muskat and Morgan values are based on an approximate solution expanded in powers of b valid for less than 0.6. It will be noticed that there is good agreement. It is interesting to note that all finite journal bearings have load capacities less than that for the infinite bearing, as may be expected. This is especially true for short bearings operating at low eccentricities. But as b approaches 1, all finite-bearing load capacities approach the load capacity of the infinite bearing.

To aid calculations for a specific journal-bearing application, values of α for $0 < b < 1$ are given in Table 1.

Conclusions: It has been demonstrated that a simple analytical solution of Reynolds three-dimensional equation can be obtained for journal bearings with circumferential grooves operating with a continuous oil film. The method employed makes possible the analytical solution of other journal-bearing problems that use the Sommerfeld boundary condition.

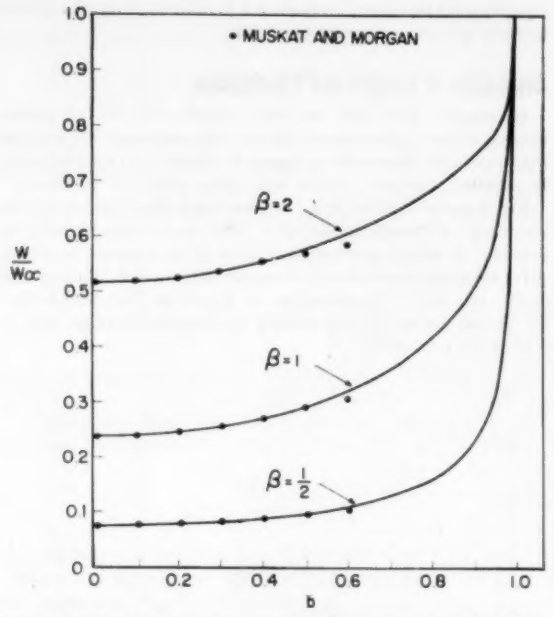


Fig. 4 Relative load-carrying capacity W/W_∞ versus eccentricity b are compared with results of M. Muskat and F. Morgan

Acknowledgment

The author is indebted to Millicent L. Fedor for her patient typing and proofreading of the manuscript. To Mr. Norman Czajkowski of the Naval Ordnance Laboratory, the author expresses sincere thanks for the aid in preparing the figures.

Table 1 Values of α versus b

b	α
0.0	1.0000
0.1	1.0057
0.2	1.0153
0.3	1.0527
0.4	1.0970
0.5	1.1594
0.6	1.2460
0.7	1.3702
0.8	1.5635
0.9	1.9372
0.95	2.3692
1.0	∞

References

1 F. Morgan and M. Muskat, "The Theory of the Thick-Film Lubrication of Flooded Journal Bearings and Bearings with Circumferential Grooves," *Journal of Applied Physics*, vol. 10, 1939, pp. 398–407.

2 M. J. Jacobson, A. Charnes, and E. Saibel, "Studies in Lubrication—X," TRANS. ASME, vol. 77, 1955, pp. 1179–1183.

3 A. Cameron and W. L. Wood, "The Full Journal Bearing," *Proceedings, I. Mech. E.*, vol. 161, 1949, pp. 59–64.

4 J. E. Meyer and F. Osterle, "On the Negative-Pressure Problem in Journal-Bearing Lubrication," doctoral thesis, Carnegie Institute of Technology, Pittsburgh, Pa., 1957.

5 H. B. Dwight, "Tables of Integrals and Other Mathematical Data," The MacMillan Company, New York, N. Y., 1949, p. 85.

6 A. E. Norton, "Lubrication," McGraw-Hill Book Company, Inc., New York, N. Y., 1949.

7 D. F. Wilcock and M. Rosenblatt, "Oil Flow, Key Factor in Sleeve-Bearing Performance," TRANS. ASME, vol. 74, 1952, pp. 849–866.

8 M. C. Shaw and E. F. Macks, "Analysis and Lubrication of Bearings," McGraw-Hill Book Company, Inc., New York, N. Y., 1949.

9 J. V. Fedor, "A Sommerfeld Solution for Finite Journal Bearings With Circumferential Grooves," doctoral thesis, University of Maryland, College Park, Md., 1958.

D. C. LEMMON
Medium Steam Turbine, Generator and Gear Department, General Electric Company, Lynn, Mass.

E. R. BOOSER
Medium A-C Motor and Generator Department, General Electric Company, Schenectady, N. Y.

Bearing Oil-Ring Performance

A series of tests was made with rings ranging from 2¾ to 16½ in. ID and at journal speeds up to 4000 fpm. Correlation of the results provides means for calculating ring speed for any journal speed. Oil viscosity, ring weight and diameter, ring immersion, and journal diameter are taken into account.

From this calculated ring speed, oil delivery can then be estimated from equations incorporating the effect of oil viscosity and related operating factors.

OIL rings have now been used for the lubrication of bearings for over a century. Because of their simplicity, reliability, and low cost, they have been installed in a great variety of rotating machinery over this period. Applied initially in relatively low-speed equipment, their range of use has continually been extended up to present limiting journal surface speeds of about 3000 to 4000 fpm.

Past studies have established many of the details of ring behavior [1, 2, 3, 4, 5].[1] The rotation of the ring depends on a propulsive force between the rotating shaft and the ring. At low speeds the static friction between the ring and journal is sufficiently high for the ring to be driven with its surface speed equal to that of the journal. At higher speeds slip occurs and the driving force is transmitted through an oil film which is analogous in many ways to that in a journal bearing. Modifications of the geometry of the inside circumference of the ring have been used in some cases to develop a higher friction coefficient and thus a more positive drive [3]. Opposing the ring rotation is the drag of the lower portion of the ring in the oil reservoir, the force required to lift the oil up from the reservoir to the top of the journal, and the frictional drag on the ring by any close-running stationary surfaces such as the sides of the ring slot in the bearing.

Oil-delivery mechanisms have also been considered. The oil lifted from the oil reservoir as a coating has been observed to be delivered to the journal largely from the inside surface of the ring by a squeezing action as the ring passes over the rotating journal [1, 2]. At low speeds some oil from the sides of the ring was found to be available, but this quantity dropped off rapidly as the centrifugal force threw out the oil at higher speeds. In the highest speed range of operation, the centrifugal force throws off as a spray most of the oil lifted from the reservoir [1]. Hersey [6] and more recently Ozdas and Ford in a comprehensive study of oil rings [5] devised dimensionless correlations of oil delivery in the full-film region of ring operation with some of the many factors involved.

These past studies have done much to provide an understanding of the fundamentals of oil-ring operation. The present analysis is directed particularly at obtaining quantitative expressions for ring speed and for oil delivery as a function of the oil characteristics, the geometry of the ring and journal, and the bearing operating conditions.

Equipment and Test Procedure

To study the performance of a number of different oil rings without the influence of the bearings which they ordinarily feed, the equipment shown in Figs. 1 and 2 was assembled. Fig. 1 is a view of the apparatus used for tests on small rings (2.75 to 6.75 in. diam), and Fig. 2 shows the arrangement used for large rings (6.75 to 16 in. diam).

In Fig. 1 can be seen a 4.50-in-diam ring running on a 2.25-in. journal. Shown are the ring guide and scraper atop the journal, the scoop and trough for collecting the oil brought up by the ring

[1] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20-22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, July 8, 1959. ASME Paper No. 59—Lub-5.

Nomenclature

d = journal diameter, in.
D = inside diameter of oil ring, in.
f = coefficient of friction
g = acceleration of gravity, 386 in/sec²
I = ring immersion, depth of ring inner diameter beneath the static oil level, in.
k = constant
k_f = full-film speed characteristic, $(DV/\nu)\left(\frac{\nu}{d\nu}\right)^{0.87}$
k_s = no-slip characteristic, $\mu V_s D/W$
n = journal rotational speed, rps
N = ring rotational speed, rps
p = length of periphery of oil-ring cross section from which oil is delivered, in. (equals inner width for unscraped ring)
Q = oil delivery, in.³/sec
Q_t = oil delivery at delivery transition speed V_t in.³/sec
v = journal surface velocity, ips
V = ring i.d. surface velocity, ips
v_0, V_0 = minimum surface velocity of journal and ring for full-film ring operation, ips
v_s, V_s = maximum surface velocity of journal and ring for no-slip operation, ips
v_t, V_t = ring oil delivery transition speeds, ips
W = weight of oil ring, lb
μ = viscosity, Reyns, lb sec/in.² (equals centipoises multiplied by 1.45×10^{-7})
ν = kinematic viscosity, in.²/sec (equals centistokes multiplied by 1.55×10^{-3})

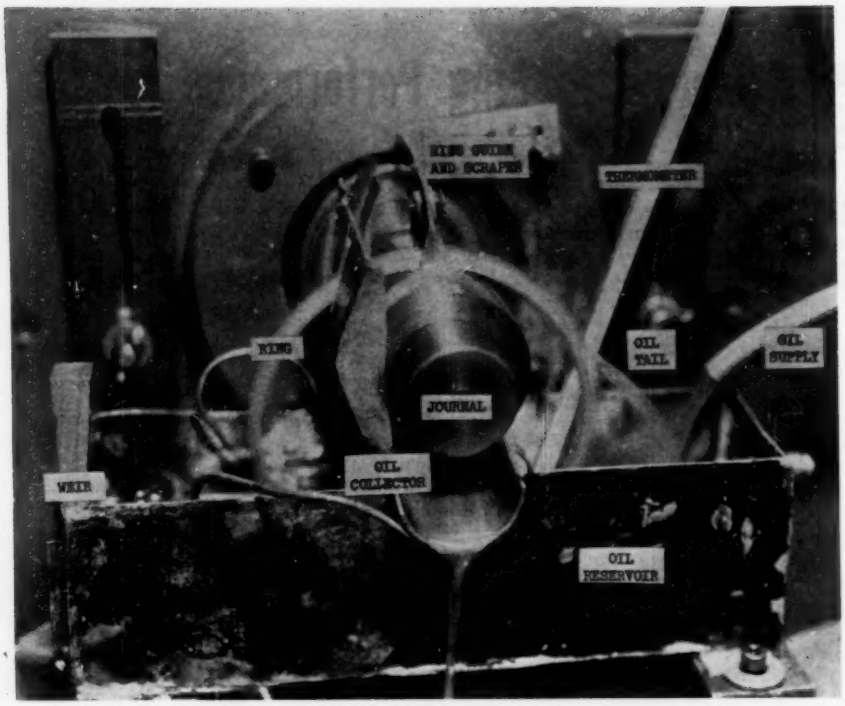


Fig. 1 Small ring apparatus

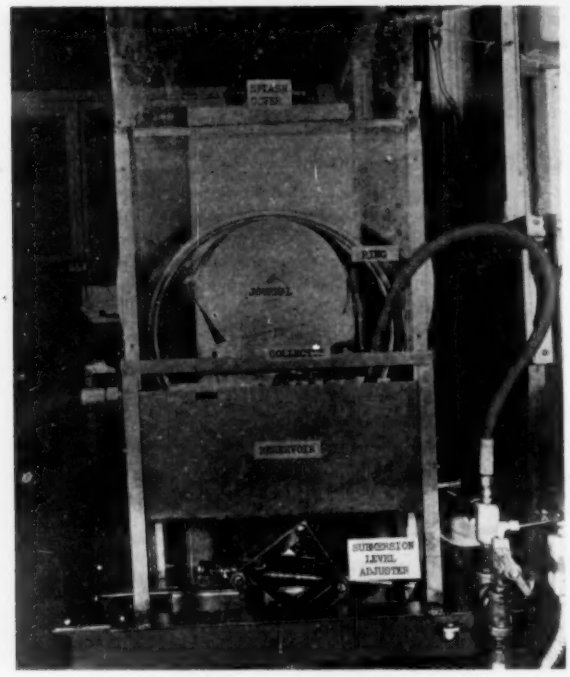


Fig. 2 Large ring apparatus

from the reservoir, the thermometer for measuring the temperature of the oil, the oil-supply pipe, and the oil tail on the ring which always accompanies operation of oil rings. The location of the weir for controlling the level of the oil in the reservoir is indicated, but the weir itself is not visible. An oil-supply pump and sump tank were also used for controlling the oil level and temperature.

In Fig. 2 a 16-in. oil ring is running on a 10-in. journal in the large ring apparatus. Visible are the oil collector, the splash cover, the reservoir, and the scissors jack for adjusting the oil reservoir and level to the desired submersion on the ring.

The shapes and dimensions of the test rings are indicated in Fig. 3. Numerical dimensions of the rings are given in Table 1. With the exception of the specially machined laminated-phenolic ring, all the rings were taken from production stock for electric motors. The trapezoidal-section zinc rings were die castings, and the T-section brass rings were made from rolled stock riveted together as shown in Fig. 3. The diameters of the steel journals run with each ring are also given in Table 1.

Table 2 summarizes the physical properties of the lubricant used in the studies, a medium turbine oil.

After the chosen journal-and-ring combination had been mounted in the apparatus with the oil collectors and ring guides in place, the oil temperature was stabilized and the ring-submersion level was adjusted. Tests were usually started at the lowest journal speed, 40 rpm. The oil issuing from the collector was directed into a graduated cylinder for an integral number of minutes to determine oil-flow rate, and the rpm of the ring was determined by counting the passages of a white mark on the ring. The speed of the journal was measured by a stroboscope. The journal speed was then increased stepwise by means of a variable drive, with a repetition of the oil-collecting and speed-measuring procedure, to 1800 or 3600 rpm or until the oil tail of the ring exceeded the boundaries of the reservoir. A few tests on small rings were run to 7000 journal rpm.

In the tests on small rings, three immersion levels were studied for each ring-and-journal combination. Tests were run at three temperatures for one immersion level for each combination. Tests at the lowest temperature, 100 F, were run first; then the oil would be heated next to 140 F and then 170 F for the subsequent series of tests. The maximum attainable temperature in

Table 1 Ring dimensions and combinations

Ring	Inner diam, D	Inner width, b	Height, c	Outer width, a	Thickness, t	Section radius, r	Material	Weight, lb	Journal diameters, d
1	2.75	0.187	0.187	0.125	...	...	Zinc	0.053	1.50
2	4.50	0.25	0.187	0.156	...	...	Zinc	0.131	1.50, 2.25, 3.00
3	4.50	0.25	0.187	0.156	...	...	Phenolic	0.026	2.25
4	5.625	0.312	0.25	0.187	...	...	Zinc	0.248	2.25, 3.00
5	5.625	0.438	0.312	...	0.094	2.00	Brass	0.41	3.00
6	6.75	0.625	0.406	...	0.094	2.00	Brass	0.66	3.00
7	10.00	0.75	0.531	...	0.141	2.00	Brass	1.75	6.00
8	16.50	1.00	0.750	...	0.234	2.00	Brass	5.00	10.00

All dimensions in inches. See Fig. 3 for symbols used.

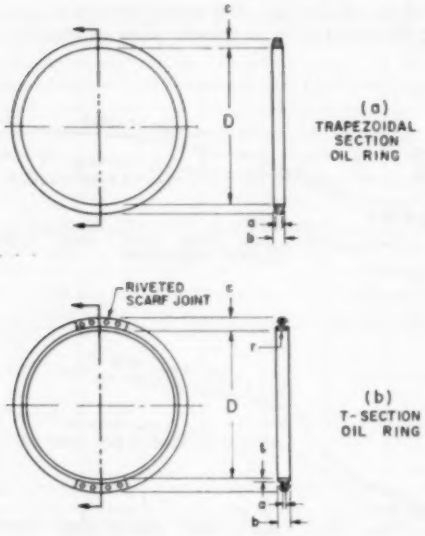


Fig. 3 Oil ring shapes

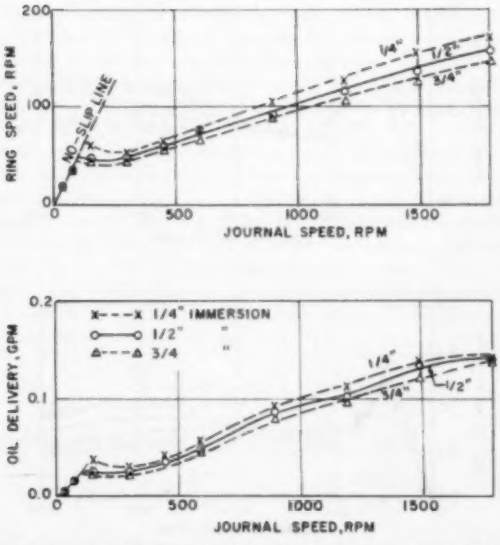


Fig. 4 Effect of depth of immersion (4 1/2-in. diam zinc trapezoidal section ring, 2 1/4-in. diam journal, 140-F oil temp)

Table 2 Oil properties

Temp., deg F	Density, gm/cc	Viscosity—Kinematic—centistokes	Viscosity—Kinematic—in.²/sec	Viscosity—Absolute, lb sec/in.²
100	0.870	73.7	0.114	9.30×10^{-6}
140	0.855	27.8	0.0431	3.45×10^{-6}
160	0.849	19.0	0.0294	2.34×10^{-6}
170	0.843	16.0	0.0248	1.96×10^{-6}

the large ring tests was 160 F because of the greater volume of oil circulating through the system.

Factors Affecting Oil Delivery

One of the requisites for satisfactory operation of a fluid-film bearing is an adequate lubricant supply. For the oil-ring bearing a vital question is always "Will the ring deliver enough oil?" How much is "enough" may be estimated by recently developed methods of hydrodynamic analysis [7, 8]. The amount of oil delivered by a ring has been observed for a wide variety of conditions in this study. Several specific factors influencing delivery will be reviewed first; then a generalized correlation of these observations will be considered in later sections.

Depth of Immersion. Fig. 4 illustrates the decrease in ring speed and decrease in oil delivery invariably found here and in previous studies [1, 3, 5] to accompany an increase in submersion. So long as the ring is sufficiently below the level of the oil in the reservoir that the inner surface is never exposed at any speed, oil delivery is not drastically changed with reasonable changes in oil level. For the variation in level experienced in bearings in actual service, i.e., from a minimum immersion of about 10 per cent of ring diameter to the maximum immersion of about 20 per cent, the change in oil delivery found in these tests was always less than 20 per cent and generally about 10 per cent or less.

Oil Viscosity. The effect of changes in oil viscosity on ring speed and oil delivery has also been extensively studied [1, 2, 3, 4, 5]. As the temperature of the oil increases, its viscosity decreases, and the quantity of lubricant delivered by the ring decreases accordingly. Data illustrating this phenomenon are plotted in Fig. 5. The viscosity of the oil affects the speed of the ring to a much lesser extent than it affects the oil delivery. This will be discussed in more detail in subsequent sections where empirical equations relating oil delivery and ring speed to viscosity are developed.

Guides and Scrapers. Perhaps the most striking observation made was the severalfold increase in oil delivery with the use of guides and scrapers to remove the oil from the sides and top of the ring and to keep the ring centered in the ring slot of a bearing. In separate tests on a complete bearing assembly, a better than 50 per cent increase in ring speed was realized when a bent paper clip was inserted between the ring and the sides of the ring slot, eliminating the long viscous drag length normally encountered whenever the ring touched either side. The use of such a guide more than doubled the oil made available to the bearing, first

because of higher ring speed and second because of the scraping action of the guide on the sides of the oil ring. Although a scraper slightly reduces the speed of a free-running ring as shown in Fig. 6, the oil delivery increases by a factor of three or four. Correlation of the data shows the amount of oil delivered to be a direct function of the length of the ring cross-section periphery from which the oil has been scraped or squeezed. The clearance between ring and scraper should be as small as possible and still permit unimpaired operation of the ring at all speeds. Theoretically the clearance may be optimized for maximum delivery, but for the rings studied this optimum was less than the 0.010 to 0.015-in. clearance required for unimpaired operation.

Material. Another variable affecting the operation of the ring is the material from which it is made. Fig. 7 gives comparative ring speeds and oil deliveries for a die-cast zinc ring and for a ring machined from a laminated phenolic plastic. Both the ring speed and oil delivery for the plastic ring are less than those for the zinc ring. Also the equilibrium position of the light ring shifts to such an extent that the ring could rub on the bearing liner or on the housing in an actual installation. All past investigators agree that lightweight rings do not operate satisfactorily. One comments [1], "Light rings, sometimes advocated by designers, have no justification."

Out-of-Roundness. Design tolerances on out-of-roundness are generally quite stringent. Surprisingly, however, the oil delivered and the ring speed for free-running rings were only slightly affected by a considerable amount of ellipticity as shown in Fig. 8. A 16.5-in-diam ring 1.0 in. out-of-round (an ellipse with a major axis of 17.0 in. and a minor axis of 16.0 in.) exhibited ring speeds and oil deliveries 80 per cent of those for a ring only 0.012 in. out-of-round. Out-of-roundness up to 2 per cent of the ring diameter has no appreciable effect on ring speed or oil

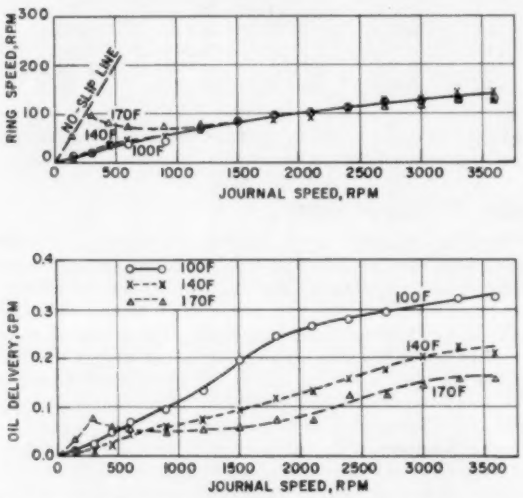


Fig. 5 Effect of temperature (5 5/8-in. diam trapezoidal section zinc ring, 2 1/4-in. diam journal, 7/8-in. immersion)

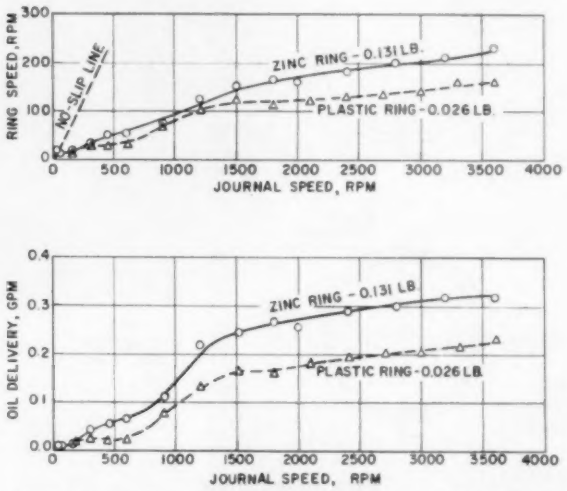


Fig. 7 Effect of ring material (4 1/2-in. diam trapezoidal section zinc ring, 2 1/4-in. diam journal, 1/2-in. immersion, 100-F oil temp)

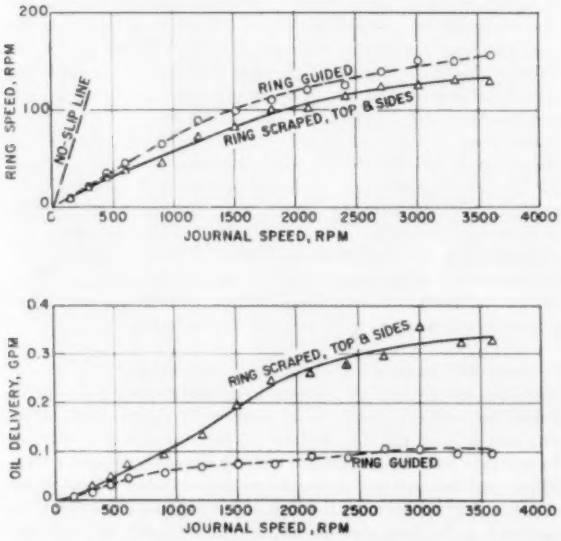


Fig. 6 Effect of scrapers (5 5/8-in. diam trapezoidal section zinc ring, 2 1/4-in. diam journal, 7/8-in. immersion, 100-F oil temp)

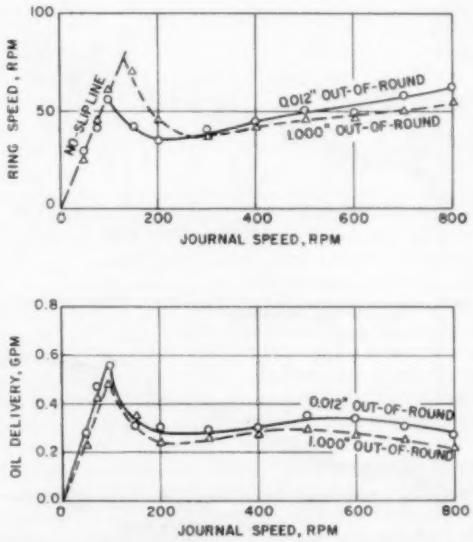


Fig. 8 Effect of ring out-of-roundness (16 1/2-in. diam T section brass ring, 10-in. diam journal, 3 1/2-in. immersion, 100-F oil temp)

delivery, but the larger ellipticities do make the rings swing in a manner which could impair their operation in a bearing. The out-of-roundness tolerances on rings should be established by the limit of this oscillation which can be tolerated in the bearing housing.

Oil-Delivery Ring-Speed Relationship

The most important factor affecting the oil delivery of an oil ring is the speed of the ring itself. As the speed of the ring is increased, the quantity of oil delivered increases—up to a point. When the velocity of the ring becomes too large, centrifugal force flings the oil off the ring and it is lost for lubrication purposes. Below this velocity, however, the relationships between oil delivery and ring speed are remarkably consistent.

Low Ring Speed. In Fig. 9 the oil delivered per length of ring cross-section periphery from which the oil is scraped or squeezed is plotted against ring velocity for a variety of rings with diameters from 2.75 to 16.50 in. with trapezoidal and T cross sections, scraped and unscraped, and with various immersion levels. With the exception of a few renegade points, the correlation is surprisingly good below the transition speed. The individual points plotted in Fig. 9 and the solid line are for an oil temperature of 100 F. Also drawn on the plot are broken lines representing the results of tests at 140 and 170-F oil temperature. The slope of all these lines on the log-log plot is the same, namely, 1.5. The intercepts are a function of temperature as it affects the lubricant viscosity. From Fig. 9 the following empirical expression is obtained for specific oil delivery as a function of ring velocity V and oil viscosity ν

$$\frac{Q}{p} = 0.026\nu^{0.68}V^{1.5}$$

$$Q = 0.14p\nu^{0.68}(DN)^{1.5} \qquad (1)$$

where Q is the cubic inches of oil delivered per second. For an unscraped ring, p becomes simply the width of the ring riding on the journal. Although this equation holds for essentially all of the tests made and is not influenced by ring immersion, several factors are not included which exerted an influence. At the same ring speed a lighter weight phenolic ring gave lower oil delivery than a zinc ring. Heavier ring weight apparently gave a thinner oil film between the ring and the shaft with more oil consequently being squeezed out from under the ring for delivery to a bearing. Surface tension of the oil probably plays a role also, but it was not evaluated here. Introduction of some such variables seems necessary to give a complete correlation of factors in Eq. (1) on a dimensionless basis.

Transition Speed. A transition ring speed, designated as V_t in Table 3, exists above which oil delivery no longer increases in the manner described in Eq. (1). A likely cause of this transition in delivery is the increased centrifugal action which reduces availability to the journal of oil exuded from between the ring and journal [2]. A dimensionless centrifugal-force term V_t^2/gD is shown to vary with a dimensionless term containing viscosity in Fig. 10 for the transition ring speed data of Table 3 for free-running (unscraped) rings as follows

$$\frac{V_t^2}{gD} = 1.9 \times 10^{-4}\left(\frac{Dg^{1/3}}{\nu^{2/3}}\right)^2 \qquad (2)$$

Substituting the numerical value of $g = 386$ in./sec.^{2}

$$V_t = 0.20\,\frac{D^{1.5}}{\nu^{0.67}}; \quad N_t = 0.063\,\frac{D^{0.5}}{\nu^{0.67}} \qquad (3)$$

This value of V_t is strongly influenced by the geometry of the system. The use of scrapers, for instance, approximately tripled

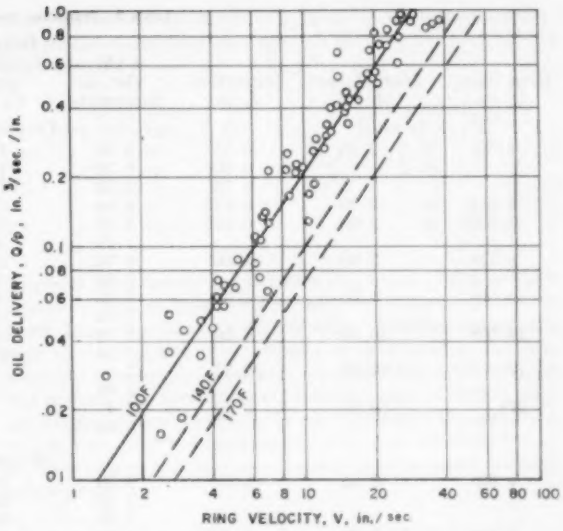


Fig. 9 Variation in oil delivery with ring speed (experimental points for 100 F only with 140 and 170 F trend lines given for comparison)

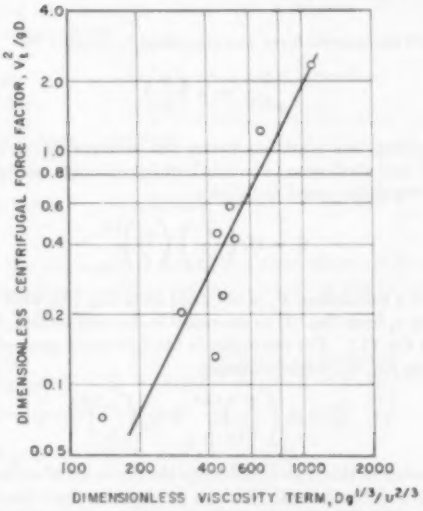


Fig. 10 Oil-delivery transition speed correlation

the values of V_t in Table 3 for the smaller ring sizes listed. A scraper still effectively removed the oil from the sides and outer ring surfaces in the volume given by Eq. (1) up to these much higher ring speeds. Eqs. (2) and (3) reflect the over-all influence of a number of factors since they are empirically obtained from experimental results. Sufficient data have not yet been obtained to separate the individual effect of each variable.

High Ring Speed. Above this transition speed, oil delivery increases much less rapidly with increasing ring speed. Surprisingly, delivery drops in this region for increasing journal speed. This was observed in a number of tests where the ring speed remained essentially constant over a wide range of journal speeds in the partial-film drive range of ring operation. Delivery varies as the square-root of the oil viscosity. The trends in the various factors involved were reasonably well related in most cases by a dimensionless oil-delivery term and a journal Reynolds number

Table 3 Transition speeds in oil-ring operation

Ring diam, D, in.	Journal diam, d, in.	Immersion, I, in.	Oil visc., μ, microreyns	Delivery transition speed, V_t, ips	No-slip limit, $v_s = V_s$, ips	No-slip charact., $k_s = \mu V_s D/W$	—Start of full-film drive— V_s, ips	v_s, ips	v_s/v_s
				Unscraped rings					
4.50	2.25	0.25	9.30	13	6	0.0019	5	17	2.8
		0.50	9.30	11	5	0.0016	4	13	2.6
		0.75	9.30	12	5	0.0016	4	14	2.8
5.625	3.00	0.875	1.96	17	23	0.0010	21	70	3.0
5.625[a]	3.00	0.875	3.45	31	24	0.0011	19	96	4.0
			1.96	36	34	0.0009	29	140	4.1
6.75[a]	3.00	1.00	9.30	22	16	0.0015	9	39	2.4
			2.34	25	33	0.0008	18	100	3.3
10.00[a]	6.00	1.00	9.30	28	38	0.0020	27	110	2.9
			2.34	69	67	0.0009	...	...	...
16.50[a]	10.00	3.50	9.30	52	72	0.0022	49	240	3.3
			2.34	121	340	0.0026	...	620+	...
17[b]	10.00	2.25	17.4	...	16	0.0011	11	42	2.6
			5.08	...	115	0.0024	44	230	2.0
27[b]	16.00	2.50	17.4	...	56	0.0016	44	142	2.5
			5.08	...	126	0.0011	51	300	2.4
				Scraped rings					
4.50	2.25	0.50	9.30	32	4	0.0013	3	10	2.5
			3.45	33	11	0.0013	10	35	3.2
			1.96	38	26	0.0017	15	54	2.1
5.625	2.25	0.875	1.96	48+	33	0.0015	22	100	3.0
Average						0.0015			2.9

[a] T-section ring.
[b] From Reference [1].

for which the general form was suggested by Hersey [6].

$$\frac{Q}{pDV} = k\left(\frac{v}{dv}\right)^{0.5} \tag{4}$$

By applying this equation twice, the delivery Q for any ring speed V and shaft speed v in this high-speed range can be related to the transition speed as follows

$$Q = Q_t\left(\frac{V}{V_t}\right)\left(\frac{v_t}{v}\right)^{0.5} \tag{5}$$

For such a calculation V_t is obtained from Eq. (3), with a corresponding v_t from Eq. (9) as discussed in the next section, and with Q_t from Eq. (1). For operation in the full-film region where Eq. (9) holds, Eq. (5) becomes simply

$$Q = Q_t\left(\frac{V}{V_t}\right)^{0.375} = Q_t\left(\frac{v}{v_t}\right)^{0.5} \tag{6}$$

At the top of this high-speed range the throw-off of oil from both the ring and the journal becomes so vigorous that the oil is no longer available for use in a bearing. Frequent and sometimes sizable deviations were encountered for the oil-delivery relations suggested for this high-ring-speed range, and further study of the factors involved would be desirable. Further analysis of the oil-delivery transition speed should be particularly fruitful.

Ring-Speed Journal-Speed Relationship

The variation in ring speed observed with variations in journal speed almost always shows three distinct regions such as can be seen in a number of the plots in Figs. 4 to 8. At low journal speeds the velocity of the inside circumference of the ring is identical to that of the OD of the journal. At some point during an increase of journal speed through this no-slip region, the static friction between the ring and journal is no longer sufficient to maintain synchronous ring speed. Beyond this journal speed a region of slower ring rotation is encountered in which the drive of the ring is partially through metallic contact and partially by a fluid film. As the journal speed goes still higher, a full-film region is finally reached in which the ring is driven entirely by hydrodynamic action through an oil film.

Tabulated in Table 3 are the transition points which were observed as the ring speed passed from one zone of operation to another with increasing journal speed. In most cases these zones of operation and their transition points were quite well defined. With smaller rings, however, the region of nonslip operation terminated at such low journal speeds that it was not always evaluated. On the other hand, with the $16^1/_2$-in. ring at 160 F the region of full-film drive was not entered throughout the test range which extended up to about 600 ips (3000 fpm).

No-Slip Operation. The ring speed in no-slip operation is quite well defined. The ring operates synchronously with the journal and the surface speeds of the two in contact are equal.

The upper speed limit for no-slip operation can be expressed through a balance of the static driving friction with the drag forces acting on the ring. The coefficient of friction between the journal and the ring multiplied by the ring weight will give the driving force on the ring which can then be visualized as being exactly equal at the no-slip limit to viscous-drag forces in the following relation

$$fW = k\,\mu V_s D$$

$$V_s = k_s\frac{W}{\mu D} \tag{7}$$

The slip characteristic k_s averages 0.0015 in Table 3 and is relatively constant with either the trapezoidal-cross-section rings or the T-section rings. From the above equation the top ring velocity for no-slip operation is seen to increase proportionately with ring weight and to decrease with increasing viscosity of the oil and diameter of the ring.

Despite the relative constancy of the slip characteristic k_s in these tests, past investigations have shown that the limit of operation without slip is also influenced by other factors. V_s can be increased severalfold by such modifications as grooving the internal surface of the ring [3] or by increasing roughness of the inside surface of the ring [5]. The slip speed also varies with the ring immersion and is higher the less the amount of drag of the ring

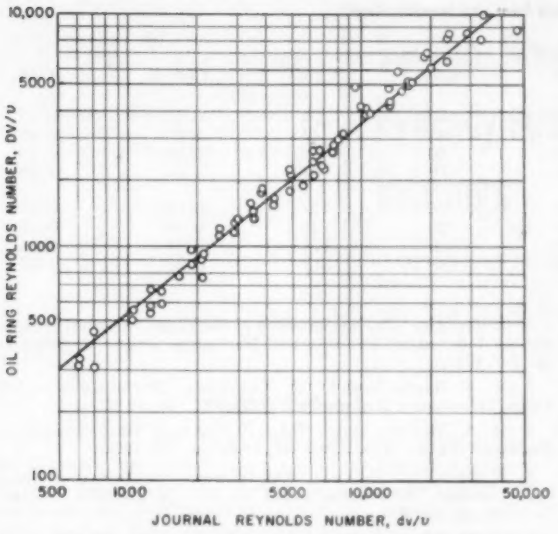


Fig. 11 Reynolds number relationship for oil ring drive

in the oil reservoir. Scraping oil from the sides and top of the ring had relatively little influence on the limit of no-slip running for the tests listed in Table 3.

Partial-Film Drive. Once the ring has started to slip with increasing journal speed, a decreased driving force is provided by the partial oil film generated between the ring and the journal. As can be seen by a comparison of V_s and V_0 in Table 3, the ring speed drops off very little with smaller rings in this region of a partial-film drive. With large rings, however, the speed decrease is quite marked. This is apparently true with large rings because of the great difference in speed between the limit in no-slip operation V_s and the full-film ring operating speed which would exist at the same journal speed.

As the degree of slipping between the ring and the journal surface increases with the increasing journal speed, a point is finally reached at which a complete lubricating-oil film is built up to support the ring. This change to full-film operation was given for the tests in Table 3 by the average relationship

$$v_0 = 2.9\, v_s \tag{8}$$

The ring-speed journal-speed relationship in this partial-film operating region was reasonably represented for most of the tests by simply connecting the points v_s, V_s and v_0, V_0 by a straight line in a logarithmic plot of ring speed versus shaft speed.

Full-Film Drive. When this region is entered, the ring speed again begins to increase with increasing journal speed. In the experimental observations, the ring speed generally increased with higher oil viscosity. Ring speed was also a factor of the diameter of both the journal and the ring. These various factors seemed satisfactorily related by comparing a Reynolds number for the driving action of the shaft with the Reynolds number for the ring in Fig. 11, which leads to the following relationship

$$\left(\frac{DV}{\nu}\right) = k_f \left(\frac{dv}{\nu}\right)^{k} = 2.10 \left(\frac{dv}{\nu}\right)^{0.80}$$

$$V = 2.10\, \frac{\nu^{0.20}}{D} (dv)^{0.80}; \quad N = 1.67\, \frac{\nu^{0.20}}{D^2} (nd^2)^{0.80} \tag{9}$$

This equation is based on an immersion of the ring in the oil in the reservoir corresponding to 15 per cent of the ring diameter. For different ring immersions, the ring speed varies inversely on an average as the 0.20 power of the ring immersion. This results in the following ring-speed characteristics for various ring immersions.

Fractional ring immersion, I/D	Ring-speed characteristic, k_f
0.05	2.62
0.10	2.28
0.15	2.10
0.20	1.98
0.25	1.90

Several other factors also influence ring speed. With scrapers present to take oil from the sides and top of the ring, the ring speed drops approximately 20 per cent. Rubbing against the sides of the ring slot in an actual ring-oiled bearing, the drag caused by a keeper placed over the top of the ring slot to keep the ring in place, and the influence of other adjacent parts in a bearing housing all serve to reduce the oil ring speed. The speed of a ring in a complete bearing assembly may drop anywhere from 5 or 10 per cent up to 50 per cent or more below the speed of a free-running ring.

Summary

The following relationships summarize the performance characteristics of oil rings observed in this study. The ring-speed shaft-speed relationships, listed for increasing journal-speed ranges, are as follows:

Range	Ring-speed equations
No-slip drive	$V = v;\ N = nd/D$
Start of slip	$V_s = 0.0015\ W/\mu D;\ N = 0.00048\ W/\mu D^2$
Partial-film drive	(See text)
Start of full-film operation	$v_0 = 2.9 v_s$
Full-film drive	$V = 2.10\, \frac{\nu^{0.2}}{D} (dv)^{0.8};\ N = 1.67\, \frac{\nu^{0.2}}{D^2} (nd^2)^{0.8}$

Once the ring speed is determined from these relations for a given operating condition, use can be made of the following oil-delivery equations.

Range	Oil-delivery equations
Low ring speed	$Q = 0.026 p_r^{0.65} V^{1.5};\ Q = 0.14 p_r^{0.65} (DN)^{1.5}$
Transition speed	$V_t = 0.20\, \frac{D^{1.5}}{p_r^{0.67}};\ N = 0.063\, \frac{D^{0.5}}{p_r^{0.67}}$
High ring speed	$Q = Q_t \left(\frac{V}{V_t}\right) \left(\frac{v_t}{v}\right)^{0.5}$

Some idea of the utility of these relations becomes apparent when they are applied to data from two previous investigations in Table 4, although the equations given were used with their units of lb, in., and sec, the comparison is provided in terms of the common rpm and gpm units. The ring speed and oil delivery were calculated from simply the shaft speed and related factors by the procedure already outlined.

For the 27-in. ring data of Karelitz [1], the oil-delivery transition ring speed was calculated from the appropriate equation to be 55 rpm. Below this speed, for 200 and 400 journal rpm, oil delivery was calculated from the calculated ring speed together with the oil-delivery equation for low ring speed. The last two cases from Karelitz are for journal speeds up to 3350 fpm, approaching the highest speeds normally encountered in ring-oiled bearing operation. For these the calculated oil delivery was obtained from the high-ring-speed oil-delivery equation.

A similar procedure was used for the 6.3-in. ring data of Ozdas and Ford [5]. Here, however, the oil-delivery transition speed was approximately 130 ring rpm as judged from the reported

Table 4 Comparison with data from past investigations

Ring diam in.	Journal diam, in.	Immersion, in.	Oil visc, centistokes	Journal rpm	Ring rpm		Oil delivery, gpm	
					Obsvd.	Calc.	Obsvd.	Calc.
From Karelitz [1]								
27	16	2.5	136	200	30	25	0.28	0.37
				400	45	43	0.85	0.86
				600	60	58	1.00	1.33
				800	70	74	1.04	1.46
From Ozdas and Ford [5]								
6.3	3.94	0.79	35.8	500	70	73	0.053	0.070
				1000	120	126	0.16	0.17
				1500	160	175	0.24	0.19
				2000	185	212	0.26	0.20

data, while the calculated value was only 66 rpm. The higher transition speed obtained from the experimental data was used together with the high-speed oil-delivery equation for the last two oil-delivery calculations in Table 4.

The ring-speed relationships held well in Table 4. Oil delivery is of the right order of magnitude, but improvement in the means for estimating the oil-delivery transition speed seems desirable. Where a higher order of accuracy is required, a few test observations with an actual bearing installation will establish more closely the numerical constants in the equations given for the system under consideration.

References

1 G. B. Karelitz, "Performance of Oil-Ring Bearings," TRANS. ASME, vol. 52, 1930, pp. 57–70.

2 E. Baildon, "The Performance of Oil Rings," General Discussion on Lubrication, Institution of Mechanical Engineers, Group I, pp. 1–7, 1937.

3 R. A. Baudry and L. M. Tichvinsky, "Performance of Oil Rings," *Mechanical Engineering*, vol. 59, 1937, pp. 89–92.

4 R. A. Baudry, "Some Thermal Effects in Oil-Ring Journal Bearings," TRANS. ASME, vol. 67, 1945, pp. 117–122.

5 N. Ozdas and H. Ford, "Oil Transfer and Cooling in Ring-Oiled Bearings," *Engineering*, August 26, 1955, pp. 268–271; October 21, 1955, pp. 570–573.

6 M. D. Hersey, "Discussion of Performance of Oil Rings," *Mechanical Engineering*, vol. 59, 1937, p. 291.

7 D. F. Wilcock, "Predicting Performance of Starved Bearings," *Lubrication Engineering*, vol. 13, 1957, pp. 348–352.

8 D. F. Wilcock and E. R. Booser, "Bearing Design and Application," McGraw-Hill Book Company, Inc., New York, N. Y., 1957.

J. S. AUSMAN
Supervisor,
Physical Analysis Unit, Autonetics,
Division of North American Aviation, Inc.,
Downey, Calif. Assoc. Mem. ASME

Torque Produced by Misalignment of Hydrodynamic Gas-Lubricated Journal Bearings

A perturbation analysis is used to find the torque generated by a misaligned journal bearing (or, equivalently, to find the angular displacement of a torque-loaded journal). Results are presented graphically in terms of nondimensional parameters.

Introduction

In some applications of gas-lubricated journal bearings, the misalignment torque or angular stiffness of the bearing is of interest. For example, a very high angular stiffness is required for gyroscope rotor bearings in order to accurately define the spin reference axis direction. Misalignment torque is also important in the design of self-aligning journal bearings. Here one is interested in determining how much friction or torsional stiffness can be tolerated in the pivoting mechanism.

An approximate analytical solution for the misalignment torque of hydrodynamic, gas-lubricated journal bearings is given in this paper. The general method follows the perturbation analysis employed in earlier papers [1 and 2].[1] Resulting characteristic curves are presented in Figs. 3 and 4. A numerical example gives a misalignment torque of 0.16 lb-in. per sec of arc misalignment. This compares favorably with similar-sized preloaded ball bearings for gyroscope rotors.

[1]Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of The American Society of Mechanical Engineers and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 3, 1959. Paper No. 59—Lub-3.

Statement of Problem

The problem is to calculate the pressure distribution and torque resulting from an angular misalignment of a gas-lubricated journal bearing. Fig. 1 shows a journal of radius r rotating at an angular velocity ω within a stationary bearing of radius $r + c$ and width (or length) b. The axes of the journal and bearing intersect at an angle δ midway between the two ends of the bearing. In other words, the journal is not translated relative to the bearing but is merely rotated or misaligned.

The bearing is operating in an atmosphere of gas at ambient pressure p_a. As the journal rotates it drags this gas through the gap h between the bearing and the journal, and sets up hydrodynamic (more properly aerodynamic or fluid dynamic) pressure forces on the journal and its bearing. For equilibrium, the integral of the pressure forces must be equal and opposite to the total load. In the case under consideration, the load is a torque of moment M. In general, this torque M will be applied at some angle ψ to the axis of the angular deflection of the journal and will have components M_1 about this axis of angular deflection and M_2 about an axis perpendicular to the angular deflection.

Since the radius of curvature r is much greater than the gap

Nomenclature

A, B, C, E = arbitrary constants
b = bearing length
c = average radial clearance between bearing and journal
d = bearing diameter
$H = \dfrac{6\eta\omega r^2}{kp_a c^2}$, dimensionless bearing parameter
h = local thickness of lubricating gas film
$i = \sqrt{-1}$
k = coefficient of polytropic expansion of gas lubricant
M = moment or torque applied to journal
p = local pressure in lubricating gas film
p_a = ambient pressure
δp_1 = first-order perturbation solution for the pressure rise, $p - p_a$, above ambient
r = radius of journal
s = shear stress
x, y, z = right-handed, orthogonal co-ordinate system with origin on the moving surface (journal) midway between the two sides of the bearing, z directed along the normal to the moving surface, and x directed in the direction of motion
$\alpha = +\sqrt{\dfrac{+\sqrt{1+H^2}+1}{2}}$
$\beta = +\sqrt{\dfrac{+\sqrt{1+H^2}-1}{2}}$
δ = misalignment angle or angular displacement of journal relative to its bearing
ϵ = eccentricity ratio
η = absolute viscosity coefficient of gas lubricant
ρ = density of gas lubricant
ζ = y/r, dimensionless axial co-ordinate with origin at midspan and directed positive in the direction of the journal spin axis
Z = a function of ζ alone
ϕ = circumferential co-ordinate angle measured positive in direction of journal rotation from line of maximum film thickness for positive values of y (from line of minimum film thickness for negative values of y)
Φ = a function of ϕ alone
ψ = angle between applied torque and the axis of the resulting angular deflection of the journal
ω = angular velocity of journal

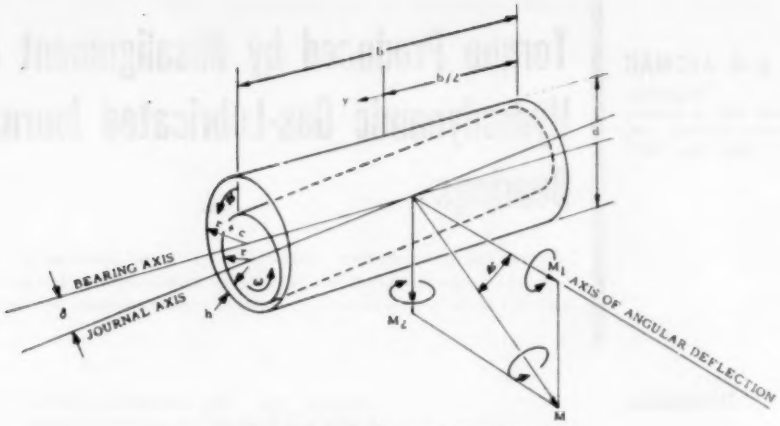


Fig. 1 Sketch of misaligned journal bearing

width h the action of the air film between the two bearing surfaces can be described by slider bearing theory. The Reynolds equation for slider bearings with lubricant density proportional to $p^{1/k}$ and constant viscosity is [1]

$$\frac{\partial}{\partial x}\left(p^{1/k}h^3\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial y}\left(p^{1/k}h^3\frac{\partial p}{\partial y}\right)=6\eta U\frac{\partial}{\partial x}(p^{1/k}h) \quad (1)$$

where p is the pressure of the lubricant, U is the velocity of the moving surface in the x-direction, η is the absolute viscosity of the gas lubricant, and k is its coefficient of polytropic expansion.

Theoretically, it is most reasonable to assume isothermal lubrication ($k = 1$), but practically it has been found that $k = \gamma$, where γ is the ratio of specific heats, often gives better agreement with experimental data [2] on journal bearings.

Equation (1) is more conveniently applied to journal bearings by the substitutions $\partial x = r\partial\phi$, $y = r\zeta$, and $U = r\omega$. With these substitutions and division by $p_0^{1/k}c^3$, Equation (1) becomes

$$\frac{\partial}{\partial\phi}\left[\left(\frac{p}{p_0}\right)^{1/k}\left(\frac{h}{c}\right)^3\frac{\partial p}{\partial\phi}\right]+\frac{\partial}{\partial\zeta}\left[\left(\frac{p}{p_0}\right)^{1/k}\left(\frac{h}{c}\right)^3\frac{\partial p}{\partial\zeta}\right]=Hkp_0\frac{\partial}{\partial\phi}\left[\left(\frac{p}{p_0}\right)^{1/k}\frac{h}{c}\right] \quad (2)$$

where

$$H=\frac{6\eta\omega r^2}{kp_0c^2}$$

Perturbation Equations

A first-order perturbation solution to Equation (2) will be obtained following the same treatment employed previously by the author [2]. In brief, the gap width is very closely approximated by

$$h = c + r\zeta\delta\cos\phi \quad (3)$$

and the pressure is assumed to be expandable in a power series of increasing order in the misalignment angle δ.

$$p = p_0 + \delta p_1 + \delta^2 p_2 + \delta^3 p_3 + \ldots \quad (4)$$

If Equation (3) and (4) are substituted into Equation (2), and terms are grouped according to powers of δ the following series of linear partial differential equation results:

$$\frac{\partial^2 p_1}{\partial\phi^2}-H\frac{\partial p_1}{\partial\phi}+\frac{\partial^2 p_1}{\partial\zeta^2}=-\frac{Hkp_0r\zeta}{c}\sin\phi \quad (5)$$

$$\frac{\partial^2 p_2}{\partial\phi^2}-H\frac{\partial p_2}{\partial\phi}+\frac{\partial^2 p_2}{\partial\zeta^2}=\frac{3}{2}\frac{Hkp_0r^2\zeta^2}{c^3}\sin 2\phi+\left[\frac{r\zeta}{c}(3\sin\phi-2H\cos\phi)-H\frac{p_1}{p_0}\right]\frac{\partial p_1}{\partial\phi}-\frac{3r}{c}\frac{\partial p_1}{\partial\zeta}\cos\phi-\frac{1}{kp_0}\left[\left(\frac{\partial p_1}{\partial\phi}\right)^2+\left(\frac{\partial p_1}{\partial\zeta}\right)^2\right] \quad (6)$$

The first equation can be solved for p_1 which can then be substituted into the right-hand side of Equation (6) permitting (6) to be solved for p_2. In theory, this procedure can be carried out to obtain higher and higher order terms in the pressure series. Practically, however, the equations rapidly become so complicated so as to force termination of the solutions after only one or two terms.

Mathematically, it is very difficult if not impossible to prove convergence of a perturbation technique such as this. However, the same method has been employed in other problems of this type, and it has been shown to give good approximations to exact numerical solutions in the case of infinitely long journal bearings [1] and to experimental results in the case of finite length journal bearings [2]. As long as the results describe the physical process to engineering accuracy, the question of convergence is academic.

At each step in the solution the pressure is required to satisfy the following boundary conditions:

1 The pressure must be periodic in ϕ; that is,

$$p(\phi) = p(\phi + 2\pi)$$

2 From dynamic and geometric symmetry, the pressure for positive ζ must equal the pressure 180 deg around the bearing at negative ζ; that is,

$$p(\phi, \zeta) = p(\phi + \pi, -\zeta)$$

3 The pressure around the two edges of the bearing must equal the ambient pressure; that is,

$$p\left(\zeta = \pm\frac{b}{d}\right) = p_a$$

Perturbation Solution

Let us try a solution for Equation (5) of the form

$$p_1 = A_1\zeta\sin\phi + A_2\zeta\cos\phi + Z(\zeta)\cdot\Phi(\phi) \quad (7)$$

where A_1 and A_2 are constants, Z is a function only of ζ, and Φ is a function only of ϕ. The form of Equation (7) is suggested by:

(*a*) The symmetry requirement for p_1; (*b*) the periodicity requirement; and (*c*) the sinusoidal term on the right-hand side of Equation (5). When Equation (7) is substituted into Equation (5), we obtain

$$Z\Phi''' - HZ\Phi'' + \Phi Z''' = \left(A_1 - HA_2 - \frac{Hkp_0\sigma}{c}\right)\zeta \sin\phi + (HA_1 + A_2)\zeta\cos\phi \quad (8)$$

In order to obtain a homogeneous, separable equation in Z and Φ, we choose A_1 and A_2 such that the right-hand side of Equation (8) is identically zero. With this choice for A_1 and A_2, Equation (7) becomes

$$p_1 = \frac{Hkp_0\sigma}{(1+H^2)c}\zeta(\sin\phi - H\cos\phi) + Z\Phi \quad (9)$$

and Equation (8) separates into the two ordinary differential equations

$$\Phi''' - H\Phi'' + K\Phi = 0 \quad (10)$$

$$Z''' - KZ = 0 \quad (11)$$

where K is the separation constant.

Solutions to Equation (10) and (11) which satisfy the first two boundary conditions (periodicity and symmetry) are

$$\Phi = Be^{i\phi} = (B_1 + iB_2)(\cos\phi + i\sin\phi) \quad (12)$$

and $$Z = Ce^{\pm\sqrt{K}\zeta} = 2C\,[\sinh(\alpha\zeta)\cos(\beta\zeta) + i\cosh(\alpha\zeta)\sin(\beta\zeta)] \quad (13)$$

where $\alpha^2 = \dfrac{\sqrt{1+H^2}+1}{2}$, $\beta^2 = \dfrac{\sqrt{1+H^2}-1}{2}$, $K = 1 + iH$, and $i = \sqrt{-1}$

Multiplying Equation (12) and (13) together and retaining either the real part or the imaginary part of the product, we obtain

$$Z\Phi = E_1[\sinh\alpha\zeta\cos\beta\zeta\cos\phi - \cosh\alpha\zeta\sin\beta\zeta\sin\phi] + E_2[\cosh\alpha\zeta\sin\beta\zeta\cos\phi + \sinh\alpha\zeta\cos\beta\zeta\sin\phi] \quad (14)$$

The constants E_1 and E_2 are fixed by the third boundary condition $\left(p_1 = 0 \text{ at } \zeta = \pm\dfrac{b}{d}\right)$. This condition is met if [see Equation (9)]

$$Z\cdot\Phi = \pm\frac{Hkp_0rb}{(1+H^2)cd}(H\cos\phi - \sin\phi) \quad (15)$$

when $\zeta = \pm\dfrac{b}{d}$

By setting $\zeta = \pm b/d$ in Equation (14) and equating the coefficients of $\sin\phi$ and of $\cos\phi$ with those in Equation (15), we can solve for E_1 and E_2. With the resulting values of E_1 and E_2, Equation (14) is then substituted into Equation (9) to obtain the first order perturbation pressure as a function of ζ and ϕ.

$$p_1 = \frac{Hkp_0b}{2c(1+H^2)}\left\{\left[\frac{d}{b}\zeta + \frac{H\cosh\left(\alpha\frac{b}{d}\right)\sin\left(\beta\frac{b}{d}\right) - \sinh\left(\alpha\frac{b}{d}\right)\cos\left(\beta\frac{b}{d}\right)}{\sinh^2\left(\alpha\frac{b}{d}\right) + \sin^2\left(\beta\frac{b}{d}\right)}\sinh(\alpha\zeta)\cos(\beta\zeta) - \frac{H\sinh\left(\alpha\frac{b}{d}\right)\cos\left(\beta\frac{b}{d}\right) + \cosh\left(\alpha\frac{b}{d}\right)\sin\left(\beta\frac{b}{d}\right)}{\sinh^2\left(\alpha\frac{b}{d}\right) + \sin^2\left(\beta\frac{b}{d}\right)}\cosh(\alpha\zeta)\sin(\beta\zeta)\right]\sin\phi + \left[\frac{H\cosh\left(\alpha\frac{b}{d}\right)\sin\left(\beta\frac{b}{d}\right) - \sinh\left(\alpha\frac{b}{d}\right)\cos\left(\beta\frac{b}{d}\right)}{\sinh^2\left(\alpha\frac{b}{d}\right) + \sin^2\left(\beta\frac{b}{d}\right)}\cosh(\alpha\zeta)\sin(\beta\zeta) - H\frac{d}{b}\zeta + \frac{H\sinh\left(\alpha\frac{b}{d}\right)\cos\left(\beta\frac{b}{d}\right) + \cosh\left(\alpha\frac{b}{d}\right)\sin\left(\beta\frac{b}{d}\right)}{\sinh^2\left(\alpha\frac{b}{d}\right) + \sin^2\left(\beta\frac{b}{d}\right)}\sinh(\alpha\zeta)\cos(\beta\zeta)\right]\cos\phi\right\}$$

where $\alpha = +\sqrt{\dfrac{+\sqrt{1+H^2}+1}{2}}$ and $\beta = +\sqrt{\dfrac{+\sqrt{1+H^2}-1}{2}}$ (16)

Two pressure distributions calculated from Equation (16) are illustrated in Fig. 2.

According to the perturbation technique outlined earlier, we should now take p_1 as given by Equation (16), substitute it into the right-hand side of Equation (6) and solve for p_2. The full solution for p_2 is quite lengthy and for our purpose it is sufficient to realize that the form of p_2 must be

$$p_2 = f_1(\zeta) + f_2(\zeta)\sin 2\phi + f_3(\zeta)\cos 2\phi \quad (17)$$

where f_1, f_2, and f_3 are functions of ζ. The important thing to observe is that p_2 contains no first harmonic terms in ϕ. As we shall see shortly, this means that p_2 contributes no support to the bearing torque load.

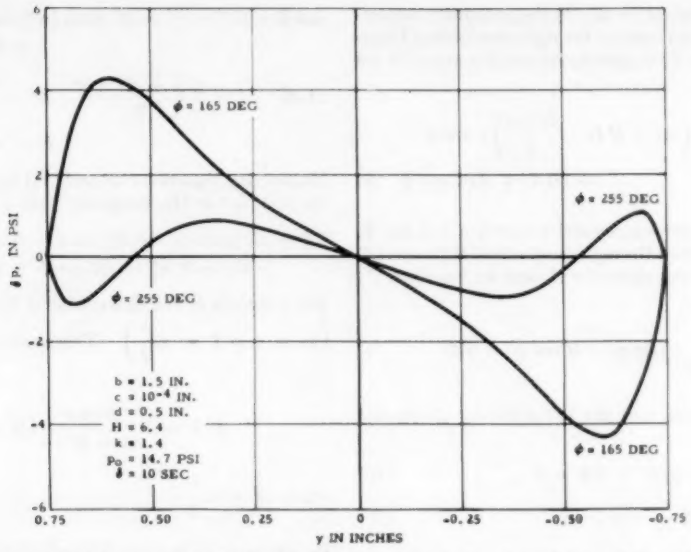


Fig. 2 Axial pressure distributions (first-order) on a misaligned gas-lubricated journal bearing

Bearing Misalignment Torque

The final step is to determine the journal restoring torque or moment M resulting from an angular displacement δ of the journal. This is also equal in magnitude but opposite in sign to the torque which when applied to the bearing would produce an angular misalignment δ of the journal relative to the bearing. The torque components can be computed from the pressure distribution by

$$M_1 = -r \int_{-\frac{b}{2}}^{+\frac{b}{2}} y\, dy \int_0^{2\pi} p \cos\phi\, d\phi \tag{18}$$

$$M_2 = r \int_{-\frac{b}{2}}^{+\frac{b}{2}} y\, dy \int_0^{2\pi} p \sin\phi\, d\phi \tag{19}$$

where M_1 is the moment component acting about the axis of angular deflection of the journal, and M_2 is the moment component acting about an axis perpendicular to the angular deflection axis. It is clear from the form of the $d\phi$ integrals in Equation (18) and (19) that only first harmonic pressure terms ($\sin\phi$ and $\cos\phi$) will contribute to the load. The first-order pressure perturbation p_1 contains first harmonics in ϕ, but the second-order perturbation p_2 does not. In general, all the odd-order perturbations p_1, p_3, p_5, etc., will contain first harmonic terms, the sum of which will determine the torque load on the journal bearing. Assuming that the predominant first harmonic terms are contained in p_1, we can determine the approximate torque load from

$$M_1 = -r^2\delta \int_{-\frac{b}{d}}^{+\frac{b}{d}} \zeta d\zeta \int_0^{2\pi} p_1 \cos\phi\, d\phi \tag{20}$$

$$M_2 = r^2\delta \int_{-\frac{b}{d}}^{+\frac{b}{d}} \zeta d\zeta \int_0^{2\pi} p_1 \sin\phi\, d\phi \tag{21}$$

Substituting Equation (16) into (20) and (21), we obtain the approximate torque components along and perpendicular to the axis of angular deflection.

$$\frac{12cM_1}{\pi k p_0 r b^2 \delta} = \frac{H}{1+H^2}\left\{H + \frac{6Hd^2}{(1+H^2)b^2} - 3\frac{(H\alpha+\beta)\sinh\left(2\alpha\frac{b}{d}\right) - (H\beta-\alpha)\sin\left(2\beta\frac{b}{d}\right)}{\frac{b}{d}\sqrt{1+H^2}\left[\cosh\left(2\alpha\frac{b}{d}\right) - \cos\left(2\beta\frac{b}{d}\right)\right]}\right\} \tag{22}$$

$$\frac{12cM_2}{\pi k p_0 r b^2 \delta} = \frac{H}{1+H^2}\left\{1 + \frac{3(1-H^2)d^2}{(1+H^2)b^2} + 3\frac{(H\beta-\alpha)\sinh\left(2\alpha\frac{b}{d}\right) + (H\alpha+\beta)\sin\left(2\beta\frac{b}{d}\right)}{\frac{b}{d}\sqrt{1+H^2}\left[\cosh\left(2\alpha\frac{b}{d}\right) - \cos\left(2\beta\frac{b}{d}\right)\right]}\right\} \tag{23}$$

Figs. 3 and 4 describe the total moment vector as calculated from Equations (22) and (23) in terms of its magnitude, $M = \sqrt{M_1^2 + M_2^2}$, and its direction, $\psi = \tan^{-1} M_2/M_1$, relative to the axis of angular deflection of the journal.

Discussion of Results

Unfortunately, there is no published experimental data on restoring torques of misaligned gas-lubricated bearings with which to compare the theoretical results just obtained. On the basis of previous work on radially loaded journal bearing [1 and 2], the author expects the first-order perturbation solutions to provide adequate engineering design data for eccentricity ratios up to one half ($\epsilon < 1/2$). In the present problem the eccentricity ratio increases linearly from zero at the center of the bearing to a maximum of $\delta b/2c$ at either end.

The same criterion (total $\epsilon < 1/2$) may also serve as a guide for the region over which the torque-loaded solution obtained herein may be superimposed on the radially loaded solution given in [2].

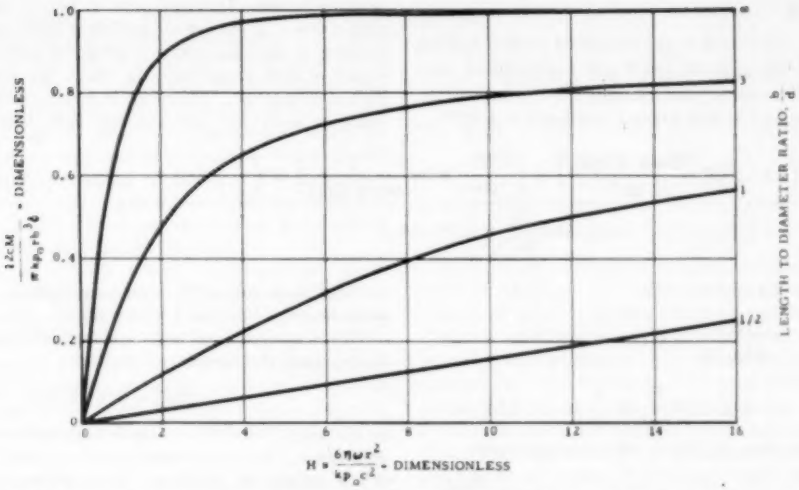


Fig. 3 Theoretical angular stiffness characteristics (first-order) of misaligned gas-lubricated journal bearings

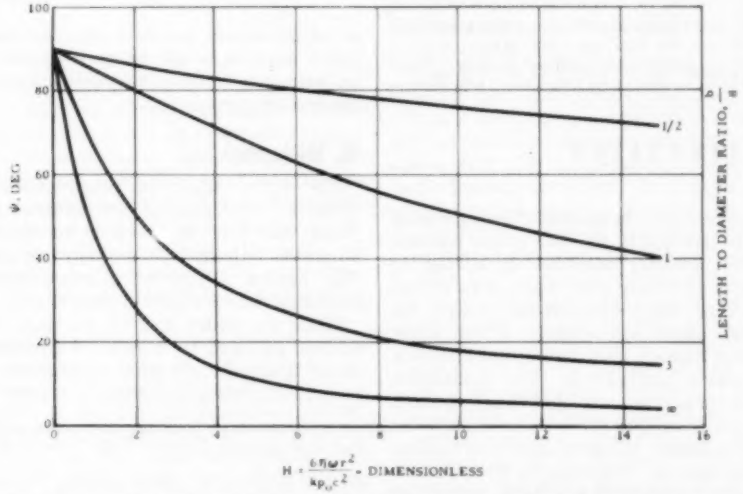


Fig. 4 Theoretical angle (first-order) between applied torque and axis of angular deflection of the journal in a gas-lubricated journal bearing

In the region of low eccentricity ratios, the solutions to the linearized (first-order) equation are fairly good. Since the equation is linear, a linear sum of the two solutions (radial-load and torque-load) is itself a solution.

The torque components, M_1 and M_2, were determined by integrating the gas film pressure distribution. It was tacitly assumed that viscous torques acting about axes perpendicular to the spin axis were negligible in comparison to the pressure derived torques. Actually, differential viscous moments must arise (Fig. 5) because viscous shear stresses are inversely proportional to the local gap width which is asymmetric when the bearing is misaligned.

The magnitude of these differential viscous stresses relative to the differential pressure stresses can be estimated as follows:

$$\frac{\text{Differential shear stress}}{\text{Differential pressure stress}} = \frac{\Delta s}{\Delta p} = \frac{\epsilon \frac{\eta\omega r}{c}}{\epsilon p_0} = \frac{Hk}{6}\frac{c}{r}$$

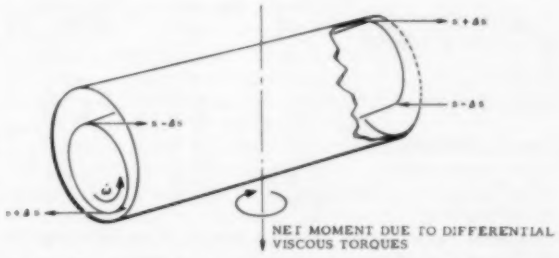


Fig. 5 Development of viscous torques perpendicular to the journal spin axis

Since $Hk/6$ is generally of the order of 1 while $c/r \ll 1$, the viscous torques perpendicular to the spin axis may be neglected in comparison to the pressure derived torques.

Numerical Example

Find the angular stiffness of a gas-lubricated journal bearing operating at 12,000 rpm in air at 100 F and 1 atmosphere pressure. The bearing diameter is $^1/_2$ in., its length is $1^1/_2$ in., and the average radial clearance between journal and bearing is 10^{-4} in.

$$H = \frac{6\eta\omega r^2}{kp_o c^2} = \frac{6 \times \left(2.8 \times 10^{-9} \frac{\text{lb-sec of time}}{\text{in.}^2}\right) \times \left(\frac{12000}{60} \times 2\pi \frac{1}{\text{sec of time}}\right) \times \left(\frac{1}{4}\right)^2 \text{in.}^2}{1.4 \times 14.7 \frac{\text{lb}}{\text{in.}^2} \times (10^{-4})^2 \text{in.}^2} = 6.4$$

From Fig. 3 for $H = 6.4$ and $b/d = 3$

$$\frac{12cM}{\pi k p_o r b^3 \delta} = 0.733$$

$$\text{or} \quad \frac{M}{\delta} = \frac{(0.733)\pi \times 1.4 \times 14.7 \frac{\text{lb}}{\text{in.}^2} \times \frac{1}{4} \text{ in.} \times (1.5)^3 \text{ in.}^3}{12 \times 10^{-4} \text{ in.} \times 2.06 \times 10^5 \text{ sec of arc/radian}}$$

$$= 0.16 \frac{\text{lb-in.}}{\text{sec of arc}}$$

References

1 J. S. Ausman, "The Fluid Dynamic Theory of Gas-Lubricated Bearings," TRANS. ASME, vol. 79, 1957, pp. 1218–1223.

2 J. S. Ausman, "Finite Gas-Lubricated Journal Bearing," Paper No. 22 presented at the IME Conference on Lubrication and Wear in London, October, 1957.

DISCUSSION

O. Pinkus[2]

It seems that a point long argued in the field of incompressible lubrication must now be renewed in the case of gas bearings. This point concerns the boundary conditions in solving the Reynolds equation. It is now fairly clear that except in rare applications, an antisymmetrical circumferential pressure distribution is an incorrect, unreal representation of the actual workings of hydrodynamic bearings using liquids as a lubricant. This is basically true also of gas bearings. The quantitative error will be much smaller in gas than in liquid bearings, but this is due not so much to compressibility effects as to the inherent low load capacity of gas bearings. But even at low unit loadings the discrepancies still seem glaring. Following are some numbers for air bearings operating in atmospheric conditions, taken from two independent sources.

Brg D × L × C	Unit loading, psi	Peak positive pressure, psig	Peak negative pressure, psig
2.5 × 3.75 × 0.00085[3]	2.1	2.2	−1.8
	3.2	5.0	−2.6
	4.3	7.5	−3.0
	5.3	10.2	−3.0
	6.4	12.6	−3.0
1 × 1 × 0.0012[4]	5.7	6.5	−3.7
	8.8	11.2	−3.7
	12.9	18.0	−3.7
	16.0	24.0	−3.7

As seen, the negative pressures are far from following the positive pressure loop and seem in fact to approach an asymptotic value. There are no published data available for bearings running at elevated ambient pressures and perhaps the author knows of such experiments in which the discrepancy is not as high as that given. However, it remains fairly certain that this behavior is typical and that the boundary conditions in gas bearings must suffer the same complications, if not worse ones, as do bearings lubricated with a liquid.

This is not intended to be a complaint against the author for having used the boundary condition

$$p(\phi) = p(\phi + 2\pi)$$

in his analysis for the writer is fully aware of the staggering difficulties in using the proper condition in an analytical solution of the differential equation. The ability to obtain analytical, albeit approximate, expressions is an achievement in itself. The point, however, is made that

$$p(\phi) = p(\phi + 2\pi)$$

is not the correct boundary condition and that if digital computers are used in solving the compressible Reynolds equation an attempt be made to formulate a more realistic condition for the negative pressure loop.

[2] Materials & Processes Laboratory, General Electric Company, Lynn, Mass. Mem. ASME.

[3] B. Sternlicht and R. C. Elwell, "Theoretical and Experimental Analysis of Hydrodynamic Gas-Lubricated Journal Bearings," TRANS. ASME, vol. 80, 1958, Fig. 12, p. 872.

[4] J. A. Cole and J. Kerr, "Observations on the Performance of Air Lubricated Bearings," Conference on Lubrication and Wear London, England, 1957, Fig. 95.3.

W. Wildmann[5]

The author should be congratulated for making another valuable contribution to hydrodynamic gas bearing literature. Those who have had occasion to design hydrodynamic gas-lubricated bearings will certainly recognize the usefulness of Figs. 3 and 4. To avoid large interpolations, a few more curves could have been included in these figures.

While the author had no published experimental data on torques produced by misaligned gas-lubricated bearings, this writer has performed some experiments on the subject. The experimental setup consisted of a rotor supported on a gas-

[5] Senior Research Engineer, Autonetics, a Division of North American Aviation, Inc., Downey, Calif. Assoc. Mem. ASME.

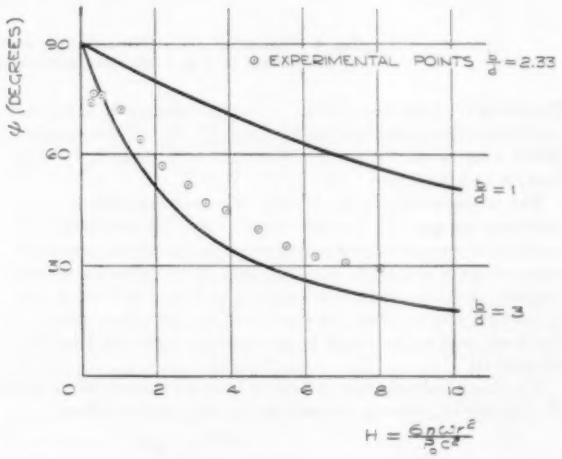


Fig. 6 Comparison between experimental and theoretical attitude angle

lubricated bearing subjected to gyroscopic torques. Measurements of attitude angle and eccentricity were made. Due to the presence of thrust pads in the experimental setup, the eccentricity measurements did not apply to the journal bearing alone and are therefore not reported. Measured attitude angles are shown in Fig. 6. It can be seen that agreement between the theory and the experimental data is very good except at low values of the dimensionless bearing parameter. This is because, at low bearing numbers, the eccentricity ratio was high and agreement with the author's perturbation solution, valid only at low eccentricity ratios, cannot be expected.

The good agreement between theory and experiment shows the validity of the paper; it also shows that agreement between theory and experiment is usually better for gas-lubricated bearings than for oil-lubricated bearings. This is surprising when one considers that complete solutions to Reynolds equations can usually be obtained for noncompressible lubrication while only approximate solutions exist for gas-lubricated journal bearings. The reason for this is the large number of assumptions concerning temperature, viscosity, and oil film breakdown that have to be made when noncompressible bearing theory is used. In gas-lubricated bearing theory, no such assumptions have to be made and the Reynolds equation with simple boundary conditions, as used by the author, gives a good description of the behavior of the lubricant film.

Author's Closure

Mr. Pinkus properly observes that peak positive (gage) pressures observed experimentally exceed the peak negative (gage) pressures in contrast to the symmetric pressure distribution predicted by first-order perturbation theory. However, the pressure periodicity condition $p(\phi) = p(\phi + 2\pi)$, cannot be blamed for this difference. Indeed, it is difficult to see how this condition could even be questioned. It states that the pressure at a point ϕ is the same as at a point $\phi + 2\pi$ which is actually the same point. If this condition were incorrect, the pressure would be a multivalued function. That is, it could have several values simultaneously at the same point.

Perhaps Mr. Pinkus is thinking about cavitation which plagues both the analysis and design of liquid-lubricated bearings. In liquid-lubricated bearings, the pressure on the negative loop may drop to the vapor pressure of the liquid but no further. At this point, part of the liquid changes state to a gas. The pressure gradient may be discontinuous under these conditions, but the periodicity requirement $p(\phi) = p(\phi + 2\pi)$ for a journal bearing still must hold true.

Of course cavitation cannot occur in a gas-lubricated journal bearing, and there is every reason to believe that under normal operating conditions the pressure distribution is not only single-valued, but also a well-behaved continuous function with continuous derivatives. Possible exceptions are:

(1) A grooved bearing with abrupt changes in film thickness which would cause abrupt changes or discontinuities in the derivative,

(2) A gas lubricant operating near its critical point in which case liquefaction might occur on the positive pressure peak and cause a discontinuous derivative, or

(3) A bearing operating with near-sonic gas velocities in which case shock waves might occur and cause an abrupt change in pressure. Even under these extreme circumstances, the pressure would still be single-valued.

What is the correct explanation for the discrepancy between experimental pressure data and first-order perturbation pressure solutions? The answer is that the first-order perturbation solution is only the first term in a series expansion of increasing order in eccentricity ratio. The series has the form

0th-order 1st-order 2nd-order

$$p = p_0 + \epsilon(F_1 \sin\phi + G_1 \cos\phi) + \epsilon^2(E_2 + F_2 \sin 2\phi + G_2 \cos 2\phi)$$

3rd-order

$$+ \epsilon^3(F_3 \sin\phi + G_3 \cos\phi + H_3 \sin 3\phi + I_3 \cos 3\phi)$$

4th-order

$$+ \epsilon^4(E_4 + F_4 \sin 2\phi + G_4 \cos 2\phi + H_4 \sin 4\phi + I_4 \cos 4\phi) + \ldots$$

E_n, F_n, G_n, H_n, and I_n are actually functions of ζ, but for purposes of discussing the circumferential pressure distribution (at a constant value of ζ) they can be considered constants.

Notice that the zeroth and first-order terms alone give a pressure which is sinusoidal with mean value p_0; the positive pressure peak is equal in magnitude to the negative pressure peak, as stated previously. If the second-order terms are included, the mean value of the pressure distribution becomes $p_0 + \epsilon^2 E_2$ instead of p_0. In effect, the whole pressure distribution is elevated. The positive pressure peak is increased while the negative pressure peak is reduced in magnitude. The resulting pressure distribution is asymmetric about p_0 in a manner similar to that observed experimentally. This then is the reason why the first-order perturbation solution is not necessarily a good approximation to the true pressure. There simply aren't enough terms included from the series.

Nevertheless, as stated in the paper, when the second-order terms are substituted into the load integrals and integrated over ϕ from 0 to 2π, the integrated effect is zero. The second-order terms do not affect the bearing load in spite of the fact that they significantly alter the pressure distribution (chiefly by raising in an algebraic sense both the positive pressure peak and the negative pressure peak). The ϵ^3 $(F_3 \sin\phi + G_3 \cos\phi)$ part of the third-order terms will contribute to the load, and these are the lowest-order terms neglected in the first-order result for the load. In other words, the first-order load expressions neglect terms of the order of ϵ^3 with respect to ϵ. In the first-order solution for pressure itself, terms of the order of ϵ^3 are neglected with respect to ϵ (if pressures are measured relative to p_0 as is the case for gage pressure). It is to be expected that the first-order perturbation solution for gage pressure will be a rather poor approximation, particularly at high eccentricity ratios.

In this regard it is well to point out that the data of Sternlicht and Elwell quoted by Mr. Pinkus in his discussion was taken at eccentricity ratios ϵ between 0.45 and 0.80. Note that at $\epsilon = 0.45$, the peak positive pressure (2.2 psig) is not greatly different from the peak negative pressure (−1.8 psig). This fact lends credence to the author's contention in the paper that he expects the first-order perturbation solutions to provide adequate engineering design data for eccentricity ratios up to one half ($\epsilon < 1/2$). Unfortunately, Cole and Kerr did not measure eccentricity. However, from a comparison of their pressure data with the data of Sternlicht and Elwell, it seems likely that the eccentricity ratio is greater than one half for all the points listed.

The author hopes that the foregoing discussion will help put the first-order perturbation solution in proper perspective for the reader and point out some of its limitations as well as its usefulness. The effect of the neglected higher-order terms is important, but it is difficult to discuss the subject adequately in the original presentation without disturbing the train of thought toward developing a solution. The author wishes to thank Mr. Pinkus for providing this additional opportunity to clarify this point.

The agreement between theory and Mr. Wildmann's experimental data on torque attitude angle is most encouraging. Note that this agreement was achieved assuming $k = 1$ which corresponds to isothermal lubrication conditions.

MILTON C. SHAW
Professor of Mechanical Engineering
Mem. ASME

ABRAHAM BER
Graduate Student of Mechanical Engineering

PIERRE A. MAMIN
Graduate Student of Mechanical Engineering,
Massachusetts Institute of Technology,
Cambridge, Mass.

Friction Characteristics of Sliding Surfaces Undergoing Subsurface Plastic Flow

It is well known that the load of an ordinary friction slider is supported by a large number of surface asperities having a collective area that is small compared with the apparent area of contact. The metal in bulk beneath such surface asperities is elastically loaded. In many metalworking operations, such as wire drawing, extruding, rolling, and metal cutting, the bulk metal undergoes plastic deformation as sliding occurs. The influence of this subsurface flow upon the coefficient of sliding friction is discussed. A simple test procedure for studying the friction characteristics of sliding metal surfaces, one of which is being subjected to plastic flow in bulk, is described, and representative data are presented for both dry and lubricated sliding.

When one metal slides over another, the normal load W is generally supported by a relatively small real area of contact A_R compared with the apparent area of contact A, Fig. 1. The characteristics of such lightly loaded sliding surfaces have been extensively studied for many years, see, for example, Reference [1].[1] The tips of the asperities of the two surfaces are believed to be plastically deformed sufficiently to provide a contact area A_R capable of supporting the load, and the friction force F is thought to result from the welding and plowing of asperities [1] or the interlocking of the slip-plane ends terminating in the asperities [2].

One of the classical laws of dry sliding surfaces is that of Amontons [3] which states that the coefficient of friction F is independent of the applied load W, where

$$f = \frac{F}{W} \tag{1}$$

Sliders of the type shown in Fig. 1 are characterized by a real area of contact that is very small compared with the apparent area of contact ($A_R/A \ll 1$). For such sliders the zone of plastic deformation is confined to the asperities and there is no plastic deformation of the metal in the subsurface in bulk. It is for situations of this sort that Amontons' law holds. Essentially all of the research that has been done on friction in recent years has been associated with such lightly loaded sliders.

There are many engineering problems where the sliding surfaces are very heavily loaded and when an attempt is made to extrapolate the results for "lightly" loaded sliders to these cases major difficulties are encountered. Heavily loaded sliding surfaces are encountered in all of the processing operations that involve heavy plastic deformation, such as: Wire drawing, extrusion, hot and cold rolling of metals; cold rolling of gears, threads, and splines; impact and explosive forming; and the cutting and grinding of metals [4]. All of these operations are characterized by the sliding of surfaces, where plastic deformation is not confined to surface asperities, but extends into the subsurface layers of the bulk metal. The laws of dry friction have not been formulated for this region and the performance characteristics of lubricants are unknown.

The relationship between the bulk plastic-flow friction regime and two adjacent bounding regions, that have been the subject of much research in the past, is shown diagrammatically in Fig. 2. Here the shear stress to which the apparent surface area A is subjected is shown plotted against the normal stress on the same area. Region I corresponds to relatively lightly loaded sliders where Amontons' law (line OB of Fig. 2) holds.

Region III, on the other hand, corresponds to the situation in ordinary materials testing where there is no free surface; i.e., where the real A_R and apparent A areas are identical. A characteristic result in materials testing is that the flow shear stress τ is independent of the normal stress on the shear plane σ, and this result is shown graphically by line CD in Fig. 2.

The region that has escaped study is the intermediate one (region II) which is characterized by the bulk plastic flow of region III, but the incomplete contact area of region I. Need

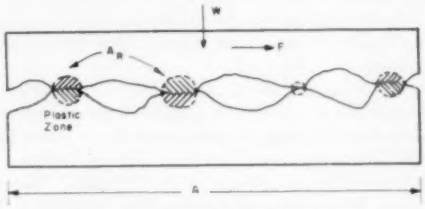


Fig. 1 Lightly loaded sliding surfaces

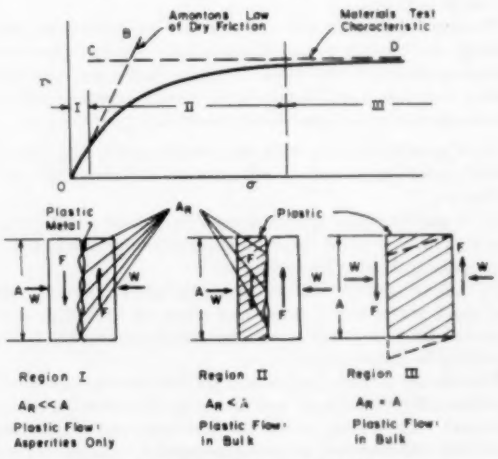


Fig. 2 The three regions of solid friction

[1] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of The American Society of Mechanical Engineers and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, July 28, 1959. This paper was not preprinted.

exists for a thorough examination of the laws of friction and of lubrication for region II and the application of these results to representative engineering processes. In this paper a test procedure for performing friction experiments in region II is described and a few preliminary results presented.

Apparatus

In considering a possible friction test, where one of the sliding surfaces is subjected to bulk plastic flow, several possibilities such as simplified wire drawing, rolling, or cutting arrangements suggest themselves. However, all of these tests are relatively complex or employ specimens of complex geometry. A very simple and convenient possibility exists in a modified Brinell hardness test. If the specimen is mounted on a ball thrust bearing as shown in Fig. 3, then the torque required to rotate the specimen against the ball may be easily measured by means of a dynamometer. The necessary bulk plastic flow is provided by making an ordinary Brinell indentation at the beginning of the test. The plastic deformation that occurs beneath a Brinell ball is quite extensive as shown by the gridded surface of Fig. 4. This specimen was produced by clamping two flat surfaces together, one of which was previously ruled with a grid, and then making a Brinell indentation at the interface of the two pieces.

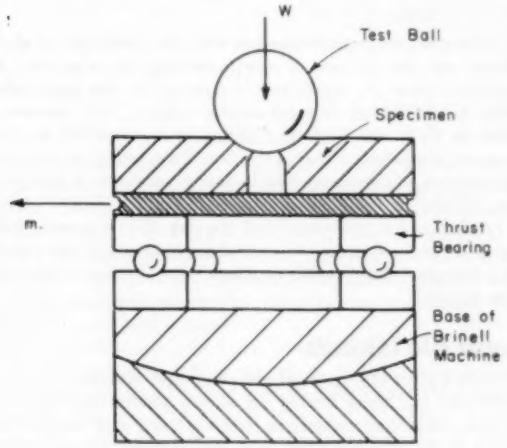


Fig. 3 Test apparatus

In order to avoid the singularity that exists at the pole in the case of an ordinary Brinell test, it was found convenient to provide a hole in the specimen at the point where the indentation was to be made. Experiments with balls and holes of different size revealed that neither of these quantities was critical, and a $^1/_2$-in-diameter ball used in conjunction with a 2-mm-diameter hole was found convenient. Ordinary bearing balls of AISI 51100 steel operating on plates of AISI 1020 steel offer an inexpensive source of specimens.

In the preliminary experiments reported here a spring balance was used to measure the torque necessary to rotate the specimen at a slow uniform speed relative to the stationary ball. The speed of rotation was without influence. Assuming the plate to be homogeneous and isotropic and the sphere to be so much harder than the plate that its deformation may be ignored, the mean friction force on the ball F was computed from the measured moment M (kg-mm) as follows

$$F = \frac{M}{\bar{r}} \tag{2}$$

where (Fig. 5)

$$\frac{\pi}{4}(d_o^2 - d_i^2)\,\bar{r} = \int_{d_i/2}^{d_o/2} r\,(2\pi r)dr = \frac{2\pi}{3}\left(\frac{d_o^3 - d_i^3}{8}\right) \tag{3}$$

and hence

$$F = 3M\left(\frac{d_o^2 - d_i^2}{d_o^3 - d_i^3}\right) \tag{4}$$

The diameter of the hole d_i after load was applied was smaller than the drilled size d_i' due to plastic flow. Fig. 6 shows a typical

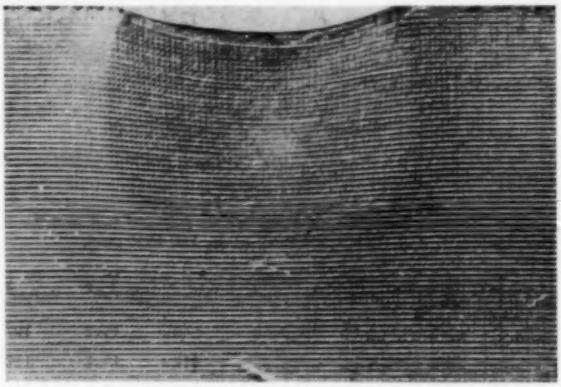


Fig. 4 Zone of plastic deformation in Brinell test. Grid was ruled on flat surfaces which were clamped tightly together and Brinell impression made at interface between the two surfaces (photograph by C. T. Yang).

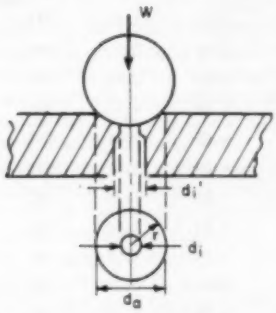


Fig. 5 Area of contact between ball and impression

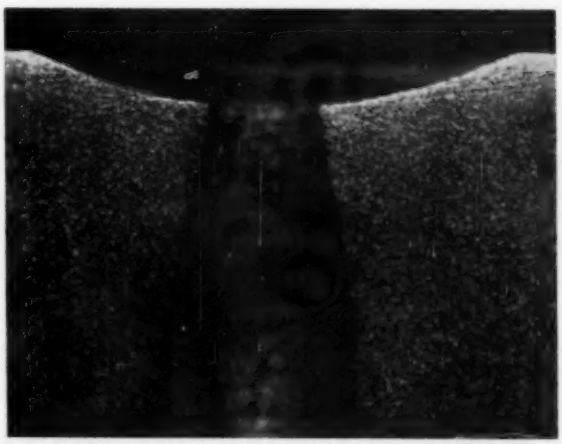


Fig. 6 Section through representative ball impression and central hole

impression produced by a $^1/_2$-in. ball loaded to 3000 kg. The diameter of the impression at the surface d_o and the diameter of the collapse hole d_i were measured in the usual manner with a Brinell microscope.

Results in Air

Results obtained using dry $^1/_2$-in-diameter balls and holes of different diameter, d_i^1, are given in Table 1. The values of mean shear stress τ and mean normal stress σ were computed as follows:

$$\tau = \frac{F}{\frac{\pi}{4}(d_o^2 - d_i^2)} = \frac{12M}{\pi(d_o^3 - d_i^3)} \quad (5)$$

$$\sigma = \frac{W}{\frac{\pi}{4}(d_o^2 - d_i^2)} \quad (6)$$

Table 1 Tests performed in air using 1/2-in. diameter ball and 2-mm hole. Materials: Ball, AISI 51,100 steel; plate, AISI 1020 steel.

d_i^1 mm	W Kg	d_o mm	d_i mm	M Kg-mm	τ Kg/mm^2	σ Kg/mm^2
4.10	1500	5.60	3.68	500	15.2	107
	2000	5.95	3.55	617	14.2	112
	2500	6.30	3.38	885	15.9	113
	3000	6.60	3.30	1065	16.1	117
	3000	6.50	3.40	1080	17.5	128
3.20	1000	4.50	2.90	239	13.5	107
	1500	5.05	2.70	440	15.2	106
	2000	5.40	2.56	640	17.2	113
	2500	5.65	2.41	785	17.9	122
	3000	6.05	2.35	886	16.1	125
	3000	6.00	2.45	970	18.3	128
2.10	1000	3.81	1.70	193	14.5	110
	1500	4.32	1.62	339	16.9	119
	2000	4.75	1.50	485	17.7	126
	2500	5.11	1.41	633	18.2	132
	3000	5.48	1.30	765	17.9	135
	3000	5.45	1.39	771	18.4	138
0	3000	5.00	0	771	18.4	152
		Tests Under Light Load				
6.35	22.7	-*	6.35	18.6	f = 0.26	
	90.7	6.50	6.35	64.8	13.4	59.8
	181.4	6.60	6.30	139.	13.2	59.6

*No plastic deformation in bulk.

The tests for increasing loads (1500 to 3000 kg) were performed upon the same hole, the load being increased to the next higher value after the torque had been measured. The duplicate test for the 3000-kg load was performed by applying the full load directly without stopping to measure the resistance to rotational motion. It is evident that while the two sets of values for the 3000-kg loads are closely similar, the values of τ and σ for the directly applied load are slightly greater in all cases. The values of τ and σ from Table 1 are shown plotted in Fig. 7.

The values given at the bottom of Table 1 were obtained using very light loads applied to a $^1/_2$-in. ball centered over a 0.250-in-diameter hole. The first load was so light that there was no evident plastic deformation at the surface of the hole at the completion of the test. The other two tests produced amounts of plastic flow that were barely visible at the edge of the hole.

The test for the lightest load, for which there was no subsurface plastic flow, yields the line marked "Amontons' law." It is seen that above a normal stress of about 30 kg/mm^2 the friction curve ceases to obey Amontons' law, and the coefficient of friction continues to decrease with increase in normal stress, σ. Lines of constant coefficient of friction are shown in Fig. 7 for reference, and it is seen that whereas the coefficient of friction is 0.26 for a very light load, it drops to a value about half as great for the heavier loads.

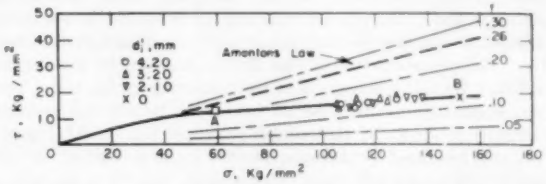


Fig. 7 Variation of τ with σ for unlubricated tests using holes of different diameter

A question frequently asked is why the coefficient of sliding friction on the surfaces of a wire-drawing die is so very low ($\cong$0.10), while the coefficient of friction for the same metals when lightly loaded is substantially higher. The answer appears to lie in the fact that sliding is accompanied by large amounts of subsurface flow in the case of a drawing operation and the presence of significant flow in bulk gives rise to a decrease in the coefficient of friction as illustrated in Fig. 7.

From Fig. 7 it is evident that the size of hole is not critical. A hole of 2-mm diameter was found convenient and was adopted for subsequent experiments in which solid and liquid lubricants were applied.

Results With Lubricants

Table 2 gives the observed and computed data for a variety of lubricants and these results are shown plotted in Fig. 8. In making these tests the central hole was filled with the lubricant and a thick layer extended over the surface. Excess solid or liquid lubricant was displaced from the test area as load was applied to the ball. The dry curve from Fig. 7 is shown dotted for reference. Those materials that yielded results below the dry curve may be considered to be good lubricants under the heavy flow conditions of this experiment. These materials include molybdenum disulfide and graphite which are two well-known solid lubricants together with ferric chloride. Ordinary mineral oil is seen to give results that are only slightly better than those for air.

Carbon tetrachloride, manganese sulfide, and finely divided aluminum oxide were found to be poor lubricants. While it is not surprising to find abrasive aluminum oxide among the poor

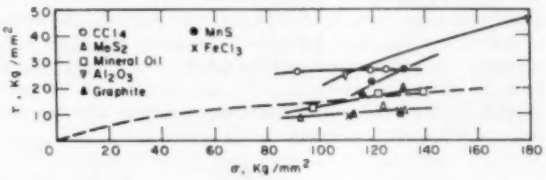


Fig. 8 Variation of τ with σ for lubricated tests

Table 2 Tests performed with lubricants. Ball diameter, 1/2 in.; hole diameter, 2 mm. Materials: Ball, AISI 51,100 steel; plate, AISI 1020 steel.

Lub.	W Kg	d_o mm	d_i mm	M Kg-mm	τ Kg/mm^2	σ Kg/mm^2
CCl_4	1000	4.06	1.60	439	26.4	92
	2000	4.80	1.40	770	27.0	120
	3000	5.55	1.22	1220	27.1	125
MoS_2	1000	3.95	1.70	124	8.3	93
	2000	4.75	1.41	354	12.9	124
	3000	5.50	1.15	479	11.2	132
Mineral Oil	1000	4.00	1.70	188	12.2	97
	2000	4.81	1.49	509	17.8	122
	3000	5.42	1.34	769	18.6	139
Al_2O_3	1000	3.90	1.90	354	25.6	110
	2000	5.20	1.40	1700	46.5	180
Graphite	1000	3.75	1.65	124	9.8	113
	2000	4.60	1.30	276	11.0	131
	3000	5.50	1.10	895	20.6	132
MnS	1000	3.80	1.85	216	16.9	116
	2000	4.82	1.40	664	22.9	120
	3000	5.60	1.45	1270	27.8	132
$FeCl_3$	1000	3.80	1.70	124	9.4	111
	2000	4.65	1.41	310	11.9	130
	3000	5.50	1.13	510	11.7	132

lubricants, it is revealing to find carbon tetrachloride and manganese sulfide yielding such poor results. In the metal-cutting literature it is frequently reported that manganese sulfide present in free-machining steels is beneficial by virtue of its ability to lubricate the chip-tool interface. This theory can be shown to be incorrect in terms of fundamental cutting-force studies and the results reported here support the view that manganese sulfide is not a good solid lubricant.

While carbon tetrachloride is in fact an effective cutting fluid at low cutting speeds, this is not due to its lubricating ability. The beneficial action of carbon tetrachloride is due to the chemical action that occurs at the tip of a cutting tool and the influence this action has upon the tendency for the chip to strain harden [5]. Although ferric chloride is formed in the cracks which propagate at the tip of the cutting tool the amount produced is exceedingly small. If substantial quantities of ferric chloride were formed at the tool tip this material should act as an effective lubricant. The fact that insufficient ferric chloride is produced at the tool tip to act as an effective lubricant when cutting in the presence of carbon tetrachloride can be demonstrated by reversing the tool at the end of a slow-speed planing cut and measuring the friction on the return stroke. When this is done the friction is found to be higher than that obtained in air, clearly demonstrating that the coefficient of friction between tool and work is that corresponding to carbon tetrachloride and not to ferric chloride.

When carbon tetrachloride is used as a wire-drawing lubricant the result is not confused by superposition of an influence on the fracture process at the tool tip and a lubricating function on the tool face. In the case of wire drawing, carbon tetrachloride can act only as a "lubricant" and in this case it is found to cause a significant increase in the coefficient of die-face friction, as Fig. 8 would suggest.

Concluding Remarks

It is shown that in general three regimes of friction exist depending upon the ratio of the real to the apparent area of contact. The simple modified Brinell test described enables data in the intermediate-friction regime to be obtained and the relation of the friction characteristics for this region to that of the conventional region of lightly loaded sliders to be compared. The simple test procedure presented should prove useful in studying the detailed action of the lubricants for use in drawing and other processes involving extensive subsurface plastic flow. It would further appear that useful wear results could be obtained from the simple system consisting of a Brinell ball rotated at constant speed in its impression.

Acknowledgment

The authors wish to acknowledge a grant from the Shell Oil Company that was used in support of this study.

References

1 F. P. Bowden and D. Tabor, "The Friction and Lubrication of Solids," Oxford University Press, London, England, 1950.

2 I. M. Feng, "Metal Transfer and Wear," *Journal of Applied Physics*, vol. 23, 1952, 1011.

3 Amontons, "Histoire de l'Academie Royale des Sciences Avec les Memoires de Mathematique et de Physique," p. 206, 1699.

4 I. Finnie and M. C. Shaw, "The Friction Process in Metal Cutting," TRANS. ASME, vol. 78, 1956, p. 1649.

5 M. C. Shaw, "On the Action of Metal Cutting Fluids at Low Speeds," *Wear*, vol. 2, 1958/59, p. 217.

DISCUSSION

Edmond E. Bisson[2]

The results reported by the authors utilizing a simple Brinell system are very interesting, particularly those on lubricated tests. The authors' results confirm the fact that graphite, MoS_2, and chloride of iron are very effective lubricants at high contact stresses. These results are, however, not surprising in view of the extensive research results obtained on this type of lubricant at very high contact stresses using a sphere sliding on a flat.

The results obtained with carbon tetrachloride are, however, quite surprising in view of the extensive mass of literature showing that CCl_4 is an effective cutting fluid at low cutting speeds. Since the generally accepted hypothesis for lubrication by CCl_4 involves reaction at the surface to form chlorides, it would seem the simplified Brinell test should, because of its low speed, give ample time for reaction. The authors' data show that ferric chloride, $FeCl_3$, is quite effective and considerable data are available in the literature to show a marked effectiveness of ferrous chloride, $FeCl_2$ references [1] and [2].[3] Since there would appear to be ample time for reaction to form chlorides of iron, a much lower friction coefficient should be expected.

I am not certain I understand the authors' statements "If substantial quantities of $FeCl_3$ were formed at the tool tip, this material should act as an effective lubricant. The fact that insufficient $FeCl_3$ is produced at the tool tip to act as an effective lubricant when cutting in the presence of CCl_4 can be demonstrated by reversing the tool at the end of a slow speed planning cut and measuring the friction on the return stroke. When this is done the friction is found to be higher than that obtained in air . . . ." It would appear to me that friction coefficient measured in such a manner (if I have deduced correctly) would always be higher than that for $FeCl_3$ because the $FeCl_3$ in most cases would be worn off immediately. Maintenance of low friction requires continual regeneration of the surface film.

[2] Assistant Chief, Fluid Systems Division, National Aeronautics and Space Administration, Lewis Research Center, Cleveland, Ohio. Mem. ASME.

[3] Numbers in brackets designate References at end of discussion.

References

1 F. P. Bowden and D. Tabor, "The Friction and Lubrication of Solids," Oxford University Press, London, England, 1950.

2 R. L. Johnson, D. Godfrey, and E. E. Bisson, "Friction of Solid Films on Steel at High Sliding Velocities," NACA TN 1578, 1948.

David Tabor[4]

This paper is valuable in focusing attention on the fact that the fundamental quantities involved in friction are the area of true contact A and the specific shear strength τ of the interface. If the conditions are such that the area of true contact is proportional to the load and independent of the size of the bodies then (assuming τ remains roughly constant) Amontons' laws of friction will be obeyed. But if the geometry or the nature of the deformation process imposes a different law on the variation of A with load and area there will necessarily be a correspondingly different law of friction. This is indeed what the authors observe.

[4] Cambridge University, Cambridge, England.

Since in their measurements the friction is usually less than $\mu = 0.2$ it is probable that little junction-growth occurs during sliding and that there is little interaction between the tangential and normal stresses. Consequently, the true critical shear strength of the interface must be close to the values deduced from the frictional measurements.

There are two points arising from this that I should like to raise:

1 Do the authors consider that the true area of contact is indeed the same as the geometric area? Even in hardness indentations this is not necessarily so.

2 There is, in general, a small increase in critical shear stress with increasing load. Do the authors think this is due to the effect of pressure on shear strength or simply to increased breakdown of surface films? The evidence rather supports the latter view since with some of the lubricant films used (e.g., graphite and MnS) there is a very large increase in shear strength with load. One wonders if this was accompanied by any marked increase in surface damage.

Machining Characteristics of Leaded Steel

FENTON L. BAGLEY, JR.
Assoc. Mem. ASME

ROY MENNELL
Rodman Laboratory, Watertown Arsenal, Watertown, Mass.

The effects of lead addition in alloy steel upon the metal-cutting process were explored over a wide range of conditions. In particular, a range of cutting speeds (from 50 to 800 fpm) and workpiece hardness (from 230 to 450 Bhn) were investigated on one workpiece material (4340) using principally a carbide (C-6) cutting tool. Orthogonal (two-dimensional) data was taken to describe the metal-cutting process, and tool-life data were obtained by running a typical production tool to failure at the various cutting conditions. Several mechanisms to explain experimental results, including lead acting as a lubricant, are discussed.

THE advantages of the addition of lead to steel have long been known, but only recently have they been measured quantitatively [3].[1] Specifically, the presence of lead has been shown to reduce cutting temperatures, increase tool life, and permit increased cutting speeds. These benefits, however, have been detected and utilized in production in areas where one or more of the following conditions occurred—low cutting speed, soft workpiece material, or thin chip [15, 21]. Little or no information has been reported on leaded steel used as a general-purpose material for high speeds and for heat-treated conditions. Some evidence was found which indicates that the advantages of lead addition decrease for hardnesses above 30 Rockwell C [22], but no information was found for high cutting speeds [500 surface feet per minute (sfpm) and above]. Therefore it was decided to investigate the effects of lead addition for a wide range of speed and a wide range of strength levels. This extension of existing data will indicate the possible future applications of leaded steels.

Design of Experiment

Although it would be desirable to determine the interactions of stock chemical composition and lead addition, the amount of testing required to obtain statistical significance for a wide range of cutting speed and material strength level would be beyond the scope of this paper. Therefore one material was used for the investigation. The material selected was 4340 and it was chosen for its favorable depth-hardening characteristics which would facilitate testing. A split heat was purchased to insure minimum chemical-composition variation between the leaded and plain material. Hardness levels of 235, 285, and 444 Bhn were used; the corresponding heat-treatments and microstructures are tabulated in the Appendix. All machining was done using C-6 carbide inserts.

Two types of information were sought in this investigation. First, an insight into the mechanism by which lead affects the cutting process, and second, an extension of machining data to higher cutting speeds and increased hardness levels. Fulfillment of the first of these objectives utilized orthogonal cutting while the second relied on full tool-life testing.

Orthogonal Cutting. To determine the effects of lead addition upon metal cutting, it is necessary to describe the process by pertinent measurements such as shear angle, machining coefficient of friction, and so forth. The fact that regular lathe turning is essentially three-dimensional, makes the task difficult, if not impossible, at the present state of the art. One difficulty is immediately evident, i.e., that shear is taking place on two opposing planes, one with a width of the depth of cut and one with a width of the feed. In addition, the nose radius complicates the shearing process as do the rake angles and the side cutting-edge angle. Under certain conditions [17] an approximation can be made using a relatively simple three-dimensional model, but the error in results may still be large and the computations tedious.

An orthogonal type of cutting was selected for describing the

[1] Numbers in brackets designate References at end of paper.

Contributed by the Lubrication Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASLE-ASME Lubrication Conference, New York, N. Y., October 20–22, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, September 25, 1959. This paper was not preprinted.

Nomenclature

a = Taylor-equation exponent of feed
A = area machined in tool-life test, sq in.
V = surface speed of material to be cut, fpm
N = number of test items
n = Taylor-equation exponent of T
C = Taylor-equation constant
r = correlation coefficient
σ = normal stress, psi
D = diagonal of Knoop indentation, Filar units
F_f = feed force, lb
F_t = tangential force, lb
r_c = chip-length ratio
$\bar{X}$ = arithmetic mean of sample
τ = shear stress, psi
γ = shear strain
R = range of sample
ϕ = shear angle, deg
μ = coefficient of friction, machining
f = feed, in. per revolution
d = depth of cut, in.
α = rake angle, side
T = tool life, min
ρ = density of the steel stock, grams/cm^3
k = thermal conductivity, cal sec per cm deg C
c = specific heat, cal per gram deg C
λ = friction angle, $\tan^{-1}\mu$
z = exponent of feed in F_t relation
b = exponent of depth in F_t relation
θ_T = tool temperature, deg C
$\Re = \dfrac{xT_d + y}{x}$
y = cost of cutting edge, cents
x = cost of machine, operator, and overhead, cents/min
ξ = Taylor-equation constant
I = Knoop hardness number
C_I = Knoop hardness constant
C_0 = constant—F_t relation
m = "constant" in temperature relation
T_d = tool changing time, min
φ' = ordinate of plasticity plot, deg

metal-cutting process in this paper because the results would be more accurate and the computational work reduced. The results, however, may not describe production turning because test conditions and tool geometry are altered to approximate the orthogonal cutting. In the present case it was felt that, although the orthogonal tests would not give a quantitative description of production turning, the tests would give valid qualitative information.

The most widely used method of obtaining orthogonal data is to use a tube of the desired material as the workpiece and a cutting tool such that its cutting edge is perpendicular to the axis of the tube and removes a chip of width equal to the tube-wall thickness. An alternate method of generating orthogonal data was to design a tool geometry that would accomplish two-dimensional cutting on solid bar stock. It was reasoned that if the side cutting-edge angle and the rake angles were zero, the nose radius very small, and the feed small with respect to the depth of cut, the result would approximate orthogonal cutting. These conditions were tested and it was found that the width of the chip was less than 0.008 in. larger than the depth of cut. This fact revealed that there was little side flow in the chip, and therefore, that the cutting mechanism was essentially two dimensional. The conditions used to obtain the orthogonal data were as follows:

BR[a]	= 0 deg	Negative land = 0.0005 in. at 45 deg to the top
SRa	= 0 deg	face of the insert
ER	= 5 deg	Feed = 0.005 in. per revolution
SRe	= 5 deg	Depth of cut = 0.200 in.
ECEA	= 5 deg	Chip breaker = $^3/_{16}$ in. from the cutting edge
SCEA	= 0 deg	(no contact with chip)
NR	= 0.20 in.	

[a] See Fig. 1 for tool nomenclature.

Because variation in tool grinding may increase the scatter in test data, special control was exercised in grinding the tools (for both orthogonal and tool-life testing). Each surface was ground with a 320-grit resinoid-bonded diamond wheel, so that the resulting surface finish averaged 8 microin. or less. The top face of all tools, however, was lapped by hand on a cast-iron block with 400-grit boron-carbide abrasive.

Preliminary tests [1] were run to determine where in the life of the orthogonal tool it is appropriate to take representative data. The results for carbide tool under the previous orthogonal conditions were as follows: (*a*) The machining coefficient of friction increases from an initial minimum value and approaches asymptotically a higher value as the tool wears; (*b*) the length of time to this steady-state condition is a function of the cutting speed; and (*c*) a single tool will never "wear in" (i.e., give a repeatable value of coefficient of friction) when it is used at intermittently varying speeds. Therefore it was concluded that the only consistent method of obtaining orthogonal carbide data for this paper was to use an initial value which was also the minimum value. After a tool had been used once, it was lapped and used again. Tools were not used more than these two times between grindings. No difference in results was detected between the freshly ground and the lapped tools. On the other hand, when a tool was lapped four or five times it gave varying results because the cumulative wear on the nose of the tool was not completely removed by lapping.

Tool-Life Testing. The full tool-life tests were conducted at varying cutting speed for each hardness, thereby furnishing information for a wide range of production applications. To facilitate usage of the data, the tool geometry chosen was considered typical for a rough turning operation with a throw-away insert-type tool, using a 0.011 feed and 0.100 depth of cut. This negative-rake tool holder is described in Fig. 1. To standardize the testing and reduce data variation the chip-breaker distance from the cutting edge was $^3/_{32}$ in.; and instead of honing, the cutting edge was chamfered with a negative land at 45 deg and with 0.001 in. width as measured from the top face of the insert.

Complete tool failure was selected as the tool-life test criterion. A previous study [8] revealed that full tool-life testing is always correct while accelerated testing techniques often incorrectly describe the effects of machining variables.

Test Equipment

All testing was performed on a 25-hp, $2^1/_2$-in. turret lathe modified for machinability testing. The lathe is driven by a variable speed d-c motor which provides an infinite speed variation up to a spindle speed of 2000 rpm. A servosystem is incorporated into the drive to maintain constant motor speed irrespective of load. The time constant of this system is 2 sec with a steady-state error of 1 per cent. This arrangement allows accurate control of the machining surface speed. Surface speed was measured with a friction-wheel hand tachometer to an accuracy of 2 per cent. Tool forces were measured by a three-dimensional strain-gage dynamometer. The tool holder was clamped into the dynamometer by set screws with positioning accomplished by the use of shim stock. A four-channel oscillograph recorder provided a continuous record of tool forces during machining.

Test Results

Because of the number and variety of data, the specific results of orthogonal cutting and tool-life testing are presented and interpreted separately in the first sections of test results. These data are followed by a discussion relating the results to the overall effects of lead addition.

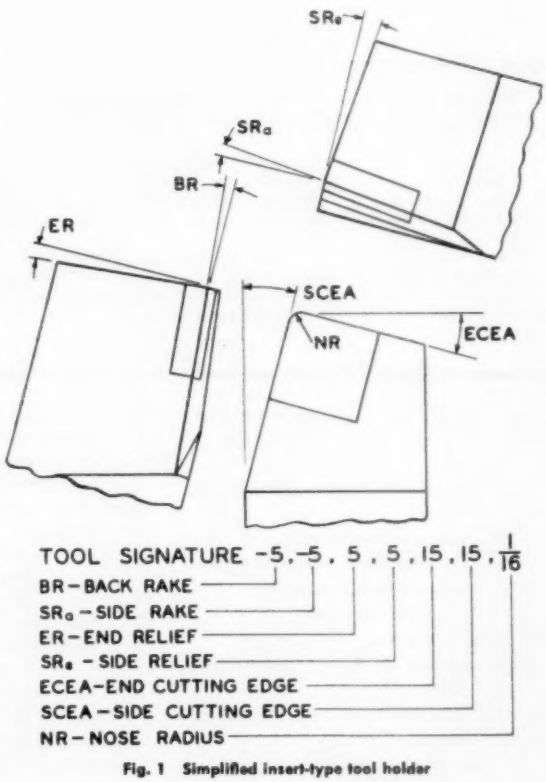


Fig. 1 Simplified insert-type tool holder

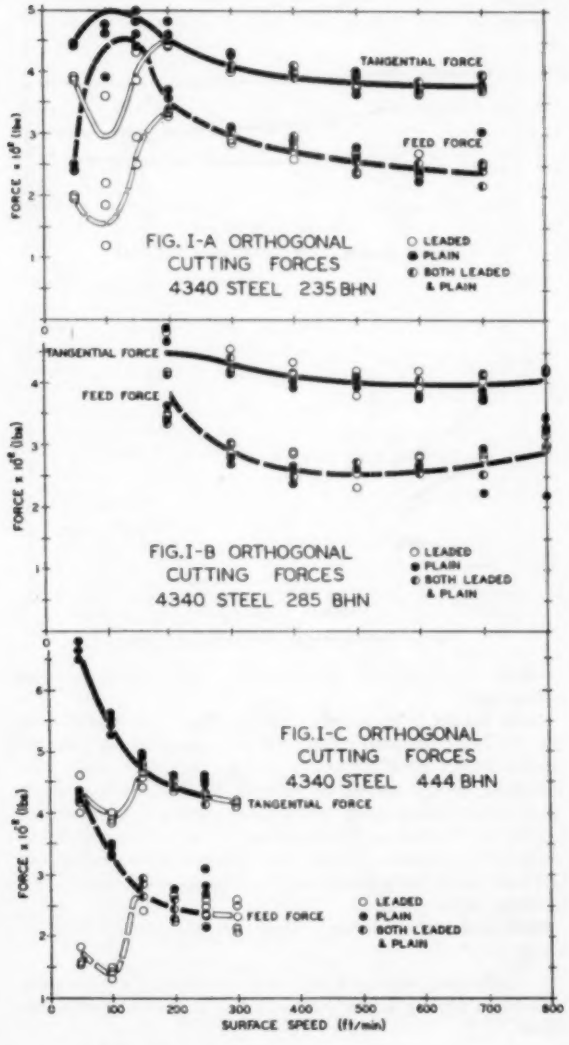
Orthogonal Cutting.

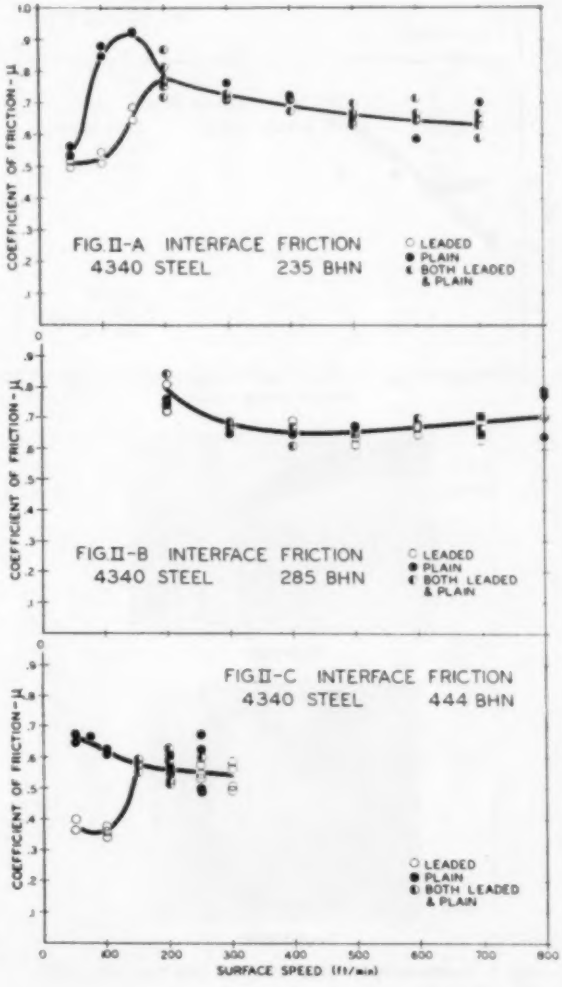
Tool Forces. Figs. I-A, I-B, and I-C show forces on the orthogonal cutting tool as a function of speed for 235, 285, and 444 Bhn, respectively. (The half-shaded circles indicate where leaded and plain-steel data coincide.) In each case, above some critical speed, both leaded and unleaded steel have the same feed and tangential forces. Below this critical speed the leaded steel forces are lower than the unleaded steel forces. As the hardness of the material increases from 235 to 444 Bhn, the critical speed decreases from 200 to 150 sfpm. These force curves for the plain steel are typical of those previously reported in machining literature. As the speed increases from a very low value the forces on the tool increase to a maximum, and as speed continues to increase the forces decrease asymptotically to some constant level.

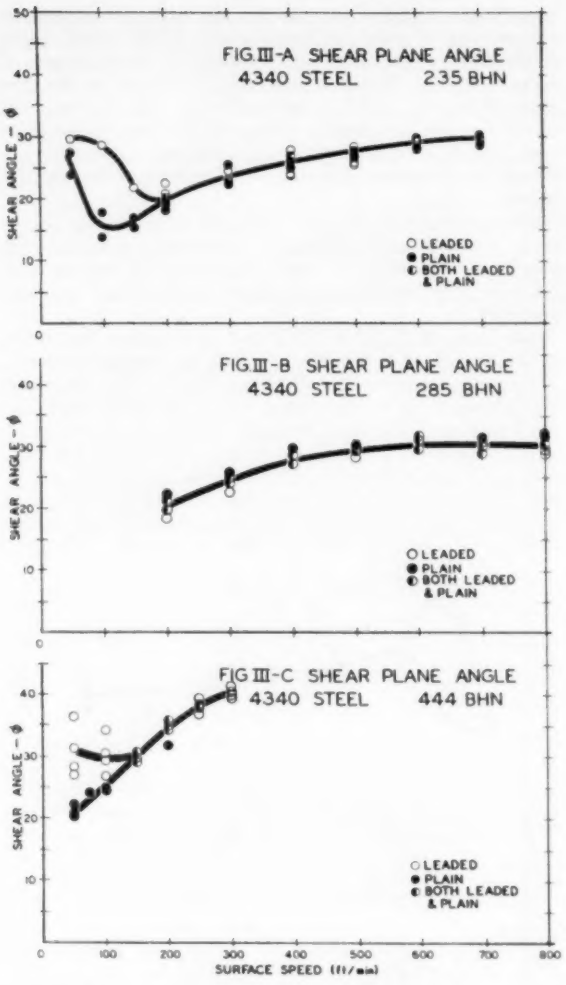
Coefficient of Friction. It may be seen from Fig. II-A that the observable effect of lead addition at 235 Bhn is below 200 sfpm where the machining coefficient of friction μ for leaded steel decreases while μ for the plain steel increases for decreasing speed and then decreases. The results at 285 Bhn, Fig. II-B, were similar to those above in the observed range of speed. Fig. II-C presents the results of the 444-Bhn test where by necessity a narrow range of speed was investigated. At the higher speeds (200 to 300 sfpm) there is a wide scatter in the results because of rapid tool wear. This wear was evidenced mostly on the nose radius of the tool. It can be seen in Fig. II-C that the wear was so rapid at 300 sfpm that no data could be taken for the plain steel. Throughout all the orthogonal testing, regardless of the cutting speed, the leaded chips showed greater curl than did the unleaded, Fig. 2.

Summarizing these results of coefficient of friction, the "beneficial" effects of lead addition, i.e., reductions of μ appear only below some critical speed which is a function of hardness and probably other machining conditions such as feed and tool geometry.

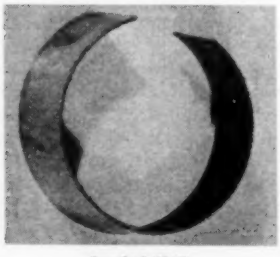
Shear Angle. Figs. III-A, III-B, and III-C contain the results of shear angle as a function of speed for varying hardness. These curves are compatible with the curves of coefficient of friction, although below the critical speed, the shear angle for the leaded steel is greater than the shear angle for the plain steel. This is to be expected. A decrease in μ is generally associated with an increase in rake angle. In contrast to the tool forces, which decrease asymptotically to a constant value with an increase in cutting speed, the shear angle increases asymptotically to a constant value as speed is increased. Both co-


TANGENTIAL FORCE
FEED FORCE
FIG. I-A ORTHOGONAL CUTTING FORCES 4340 STEEL 235 BHN
LEADED
PLAIN
BOTH LEADED & PLAIN
FIG. I-B ORTHOGONAL CUTTING FORCES 4340 STEEL 285 BHN
FIG. I-C ORTHOGONAL CUTTING FORCES 4340 STEEL 444 BHN
SURFACE SPEED (ft/min)


COEFFICIENT OF FRICTION - μ
FIG. II-A INTERFACE FRICTION 4340 STEEL 235 BHN
FIG. II-B INTERFACE FRICTION 4340 STEEL 285 BHN
FIG. II-C INTERFACE FRICTION 4340 STEEL 444 BHN
LEADED
PLAIN
BOTH LEADED & PLAIN
SURFACE SPEED (ft/min)



Plain 4340



Leaded 4340

Fig. 2 Representative orthogonal chips (235 Bhn, 400 sfpm)

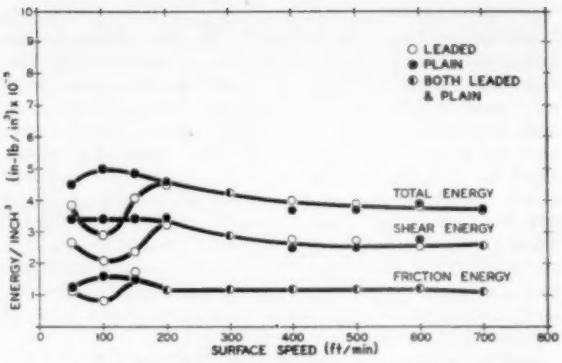


Fig. 3 Cutting energy (4340 steel, 235 Bhn)

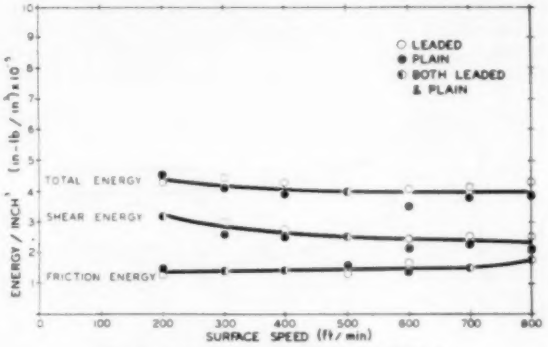


Fig. 4 Cutting energy (4340 steel, 285 Bhn)

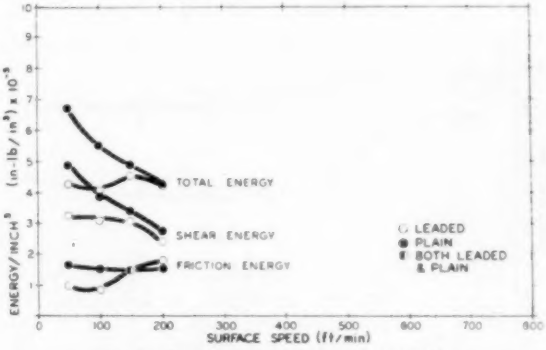


Fig. 5 Cutting energy (4340 steel, 444 Bhn)

efficient of friction and shear-angle results reveal the same critical speed.

Cutting Energies. An energy analysis (Figs. 3, 4, and 5) from the previous orthogonal cutting data displayed the fact that the friction energy remained nearly constant as speed varied while the shear energy tended to decrease as the speed increased. Above some critical speed, which again appears to be a function of hardness, there is no difference between any of the energies for leaded and plain steel. Below this critical value the leaded steel had lower shear energy and thus lower total energy than did the unleaded steel.

Results of the Orthogonal Tests. These can be summarized as follows:

1 Differences between leaded and plain steel were detected by orthogonal cutting at low speeds for all workpiece hardnesses tested.

2 The tool forces, machining coefficient of friction, shear angle, and shear energy displayed these differences below the same critical speed for given material hardness.

3 The effect of lead addition on the metal-cutting process is a function of speed and workpiece hardness.

4 The leaded-steel chips always curled more than their plain counterparts.

Tool-Life Testing. Tool-life data as a function of speed for the three hardness levels tested appears in Figs. IV-A, IV-B, and IV-C. The lines through the data were obtained by a linear regression of the log of tool life (sq in.) on the log of surface speed (fpm). The mathematical equation for each set of data appears on the curve as well as standard error estimates which include 95 per cent of the test data. For clarity the lines for these standard error estimates do not appear on the curves, but their distance from the regression lines is given in inches (on these plots one log cycle is 5 in. in length). The values of the correlation coefficient for both the leaded and the plain steel are given in the figures. The form of the equation appearing in these figures is the widely known Taylor plot and the dimensions of the equations, to agree with convention, are V in fpm, and T in minutes.

From the tool-life plots (Figs. IV-A and IV-B) it can be seen that the leaded-steel data has a greater slope. (This is shown also on the Taylor-plot equation where the exponent of T is always less for the leaded steel.) A "Student's t" test was conducted to determine whether there was a significant difference between tool life for the leaded and nonleaded data. It was concluded from this test that there was no statistically significant difference between leaded and nonleaded tool life when data were taken over the full range of speed tested for the 235 and 285-Bhn material. It should be pointed out that this does not mean there is no difference between tool life, but only that none was proved—in these tests the scatter in the data was large with respect to any difference in tool life of leaded versus nonleaded steel. An analysis of tool-life data at the upper speed did show a difference in tool life between the leaded and plain steel. At 800 sfpm tool life on the plain steel at 235 Bhn was greater to the 0.05 confidence level and at 700 sfpm tool life on the plain steel at 285 Bhn was greater to the 0.10 confidence level. On the other hand, for that part of the range of speeds tested which would be used in production (less than 600 fpm) no difference in tool life between the leaded and plain steel was proved. A similar analysis was not made for the 444-Bhn material because of the manner in which these data were obtained. At the high hardness level there was no point in the life of the tool which clearly indicated failure.

Tabulation of Knoop diagonal squared for 235 Bhn

sfpm	Position		$\bar{X}_p \times 10^{-3}$	$R_p \times 10^{-3}$	ϕ_p	$\bar{X}_L \times 10^{-3}$	$R_L \times 10^{-3}$	ϕ_L	$\lvert \bar{X}_p - \bar{X}_L \rvert$	Confidence level	N
*100	Center	∥	425	47	16°	489	50	23°	64	90%	4
100	Edge	⊥	392	58	16°	413	36	28°	21	80%	6
*100	Edge	⊥	387	47	16°	433	43	23°	46	95%	5
200	Edge	⊥	430	35	20°	409	46	20°	21	90%	6
200	Center	∥	474	90	20°	473	36	20°	1	10%	6
310	Center	∥	450	93	24°	445	61	24°	5	20%	4
310	Edge	⊥	422	59	24°	415	84	24°	7	50%	6

Knoop hardness number $\equiv I = \dfrac{L}{l^2 C_p} = \dfrac{C_I}{D^2}$

where $l = eD$; e = magnification $\quad \therefore C_I = 1924 \times 10^6$

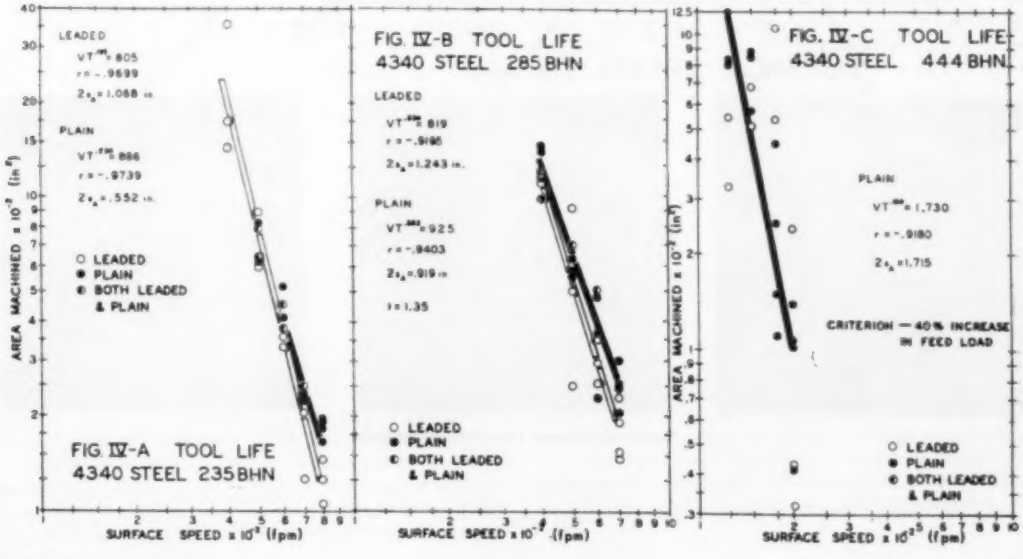
The diagonal is given in Filar units.

$\bar{X} \equiv$ arithmetic mean of sample

$R \equiv$ range of sample

$N \equiv$ number of readings in sample

The confidence level is the probability of being in error in assuming the two samples are from the same population. The statistical test was a "Student t" test using the range as an estimate of standard deviation [9].



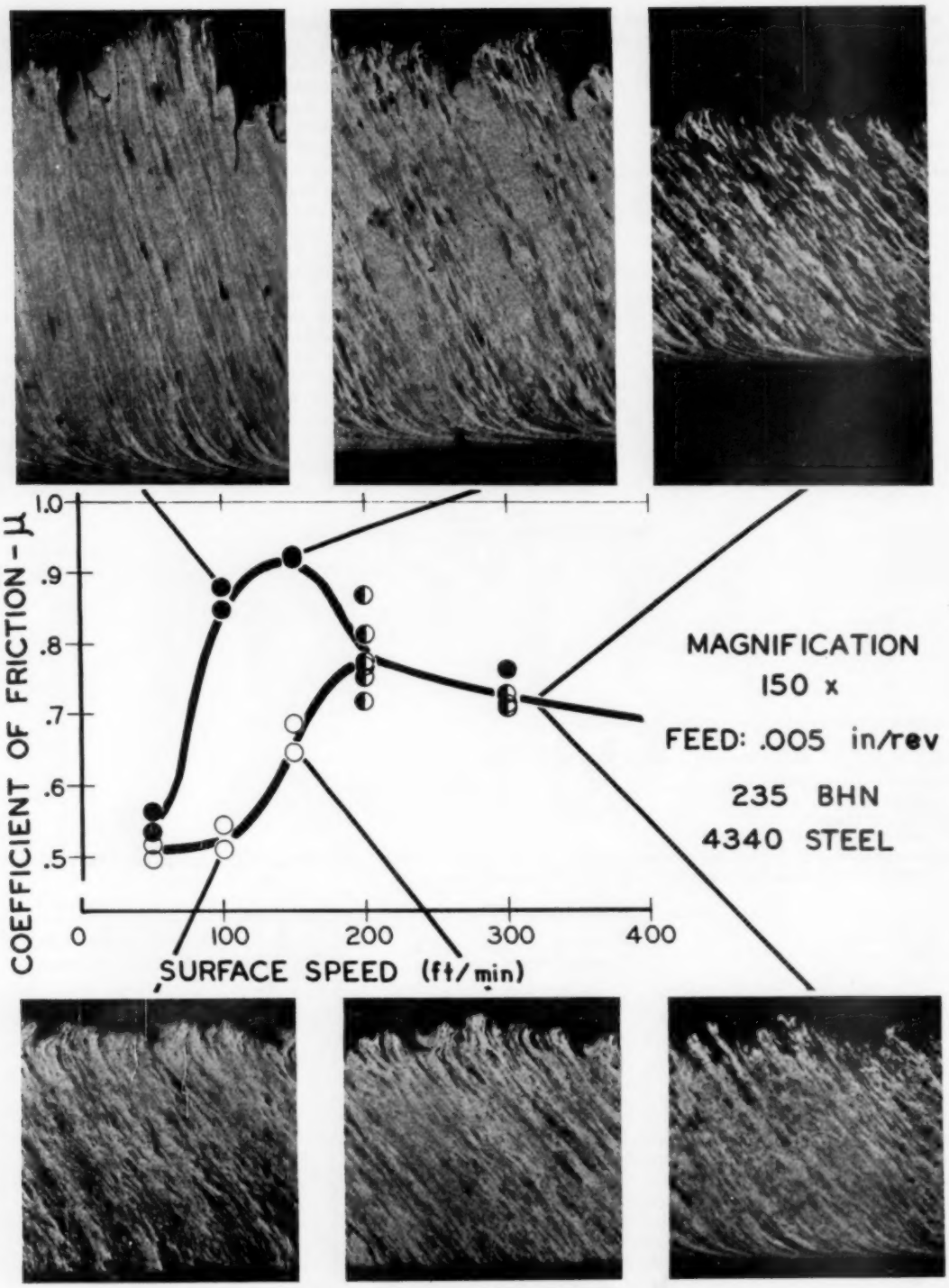


Fig. 6 Orthogonal chip photomicrographs

Tool failure resulted from increasing vibration due to excessive cratering close to the cutting edge of the tool. For convenience and because no other criteria proved suitable, 40 per cent increase in the feed force on the tool was selected as a point of failure. This definition of full tool life may partially explain the configuration of the leaded-steel data appearing in Fig. IV-C. However, chip formation of the leaded steel has a direct bearing on this curve. While the plain steel always exhibited a continuous chip, the leaded steel revealed a discontinuous chip below 200 sfpm, and a continuous chip at 200 sfpm.

Effects of Lead Addition

To augment the information directly obtainable from the orthogonal tests, several specialized investigations were performed seeking the causes of the results previously discussed.

The first was a microscopic comparison of the leaded and plain steel chips. The chips from orthogonal cutting were mounted in plastic and ground until a representative longitudinal cross section was obtained. The specimens were then etched and photomicrographs taken (see Fig. 6). An examination of Fig. 6 revealed the following effects of lead below the critical speed (200 sfpm): (*a*) Decreased chip thickness; (*b*) decreased roughness of the back of the chip; (*c*) decreased "wipe-down." The first observation is directly related to the increased shear angle produced by lead addition. The decreased roughness of the back of the chip may be a function of the shear angle. Wipe-down, as used in this paper, pertains to the curvature of the chip grain-deformation lines on the tool side of the chip caused by secondary shear at the chip-tool interface. The plain steel chip has a nearly constant wipe-down with increased speed. However, the leaded-steel chip has very little wipe-down at the lowest speed, but increased with increasing speed until above the critical speed where there is no difference in wipe-down between the leaded and plain chips.

To further explore the relationship between μ and wipe-down 1113 (leaded and plain) steel at 190 Bhn was machined using orthogonal cutting and the chip was mounted similarly to the above. The reason for the choice of 1113 was the free-machining properties of this steel (decreased μ and chip thickness) which are similar to those achieved by leading 4340. Fig. 7 displays the fact that, although μ is decreased in the plain steel (compared with 4340, Fig. 6), the wipe-down is increased. The addition of lead to this steel produced a decrease in wipe-down similar to that produced in 4340 steel, and also a decrease in μ below the critical speed ($\simeq$500 sfpm). The higher critical speed of the 1113 leaded steel should be noted. The factors involved in the position of the critical point will be discussed later in this section.

At this stage a more complete speed range was run on the 4340 steel and these chip photomicrographs are shown in Fig. 8. Fig. 8 completes the speed range of Fig. 6. Also included is a leaded-steel chip that was welded to the tool tip when the cut was interrupted.

Included in Fig. 8 is a photomicrograph of a chip after Knoop-hardness readings were taken on it. After some experimentation, the placement of the indentations, as shown in Fig. 8, was chosen to reduce variation in hardness readings. Readings in the center of the chip were taken parallel to the flow lines, while those at the edge were taken at an angle perpendicular to the center readings.

The square of the length ($D \times 10^{-3}$) of the longer diagonal of the indentations was recorded and a statistical tabulation is given.

The chip hardness varies as the inverse of D^2 tabulated above, therefore the leaded steel produces a softer chip at speeds below the critical speed. From Fig. 6 it can be seen that 100 sfpm is the point of maximum difference between leaded and plain steel at these conditions. The leaded chip of the row marked by * was produced by an orthogonal tool of negative rake ($\alpha = -5$ deg)—all others had an α of 0 deg—to obtain leaded and plain chips subjected to approximately the same shear angle. At the critical speed (200 sfpm) the leaded and plain chip hardness is similar in the center but different at the edges. As the speed further increases the hardness at the edge also tends to become similar.

Data obtained from a BUE survey (orthogonal cutting) at various feeds and speeds is found in Fig. 9. The critical speed was not sensitive to changes in feed, Fig. 9(*A*). It can be seen from Fig. 9(*B*) that the size of the BUE, average height above top face of tool, does not appear to be the controlling factor of the critical point. The contact areas recorded in Fig. 9(*C*) further document the increased chip curvature (decreased contact width) observed qualitatively in previous tests.

The following theories may offer insight into the mechanism whereby lead addition alters the metal cutting process:

1 Real versus apparent area
2 Lubrication
3 Active shear planes

1 The real-versus-apparent-area theory presented by Shaw [18] applies to the motion of the chip across the tool. It is proposed that this action is neither simple friction of an elastic body, where the real contact area is a small percentage of the apparent area, nor pure plastic flow, where sliding occurs within the material, and the apparent and real area are equal. This relationship is shown in Fig. 10(*A*). In the elastic region (1) the shearing frictional stress τ is proportional to the normal stress σ and in the plastic region (2) the shear stress is independent of the normal stress. Instead, it is proposed that motion at the tool-chip interface is a combination of elastic and plastic phenomena. An increase in shear stress so as to move from point A to point B would then decrease the machining coefficient of friction from μ_A to μ_B.

A decrease in wipe-down by the addition of lead is direct evidence that the shear stress has decreased and that the action is of a less plastic nature than without lead. However, the machining coefficient of friction decreased with the addition of lead. Therefore it must be concluded that lead changes the character of machining so that this change must be described with a different shear-normal-stress relationship. The lubrication theory suggests such a relationship.

2 The lubrication theory was developed to explain the mechanism of machining leaded steel. The argument used is that the lead smears out between the chip and the tool to form a lubricating layer. If the lead in the steel does in fact act as a lubricant, the friction force would be reduced, causing the decrease in μ that was observed below the critical speed. The decreased wipe-down of the leaded steel is in direct support of the theory. The reduction of the friction force F_f and μ related to the preceding would produce the increase in shear angle and decrease in chip thickness that were observed when machining leaded steel. On the shear-normal stress diagram, Fig. 10(*B*), this reduction in real coefficient of friction by lead addition can be portrayed by a curve lower than the first but asymptotic to the same shear stress for plastic flow. The addition of lead then would cause a movement from point B to point B′ with the result of lower shear stress and machining coefficient of friction.

3 The increased-number-of-active-slip-planes theory suggested by Rebinder and used to explain the effects of CCl_4 by Shaw [19] could be present in the machining of leaded steel. If the finely dispersed lead particles were to act as stress concentrations the number of active slip planes could be increased and the shear strain on each plane reduced, thereby producing a softer more homogeneous chip. The decreased roughness of the back of the leaded-steel chip could be the direct result of this mechanism. It is difficult to assess the value of this theory because the principal cause of the decreased hardness for leaded steel is the increase in the shear angle previously discussed in connection with 4340 steel.

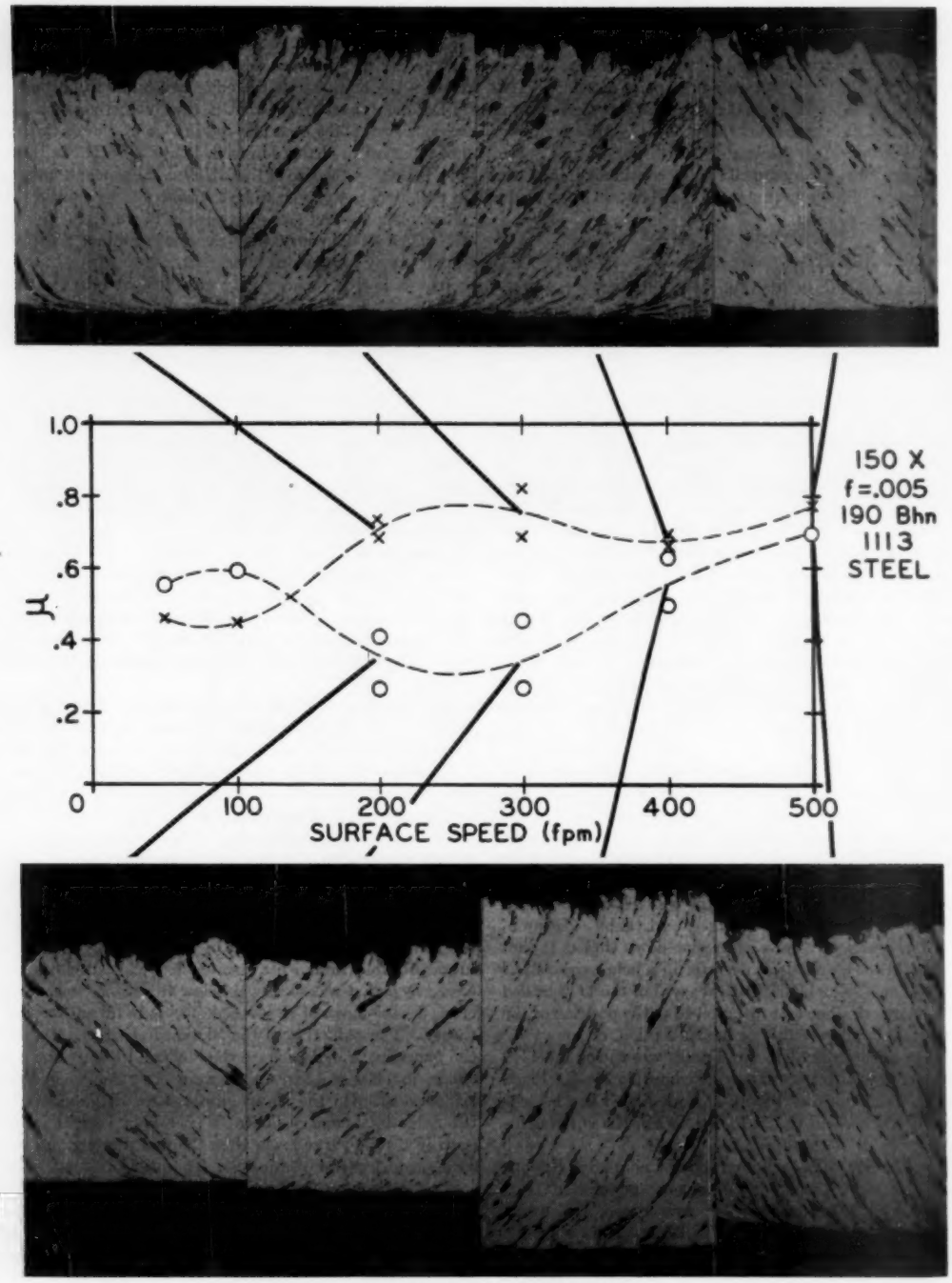


Fig. 7 Orthogonal chip photomicrographs

If it is assumed that the strain-hardening properties of the leaded and plain steels will be equal, then the comparative chip hardness will be a function of γ only. For orthogonal cutting the following relationship exists:

$$\gamma = \frac{\cos \alpha}{\sin \phi \cos (\phi - \alpha)}$$

Evaluation of γ in this case presents a problem because the relation is not strictly valid in the region of BUE which is also the region of the influence of lead addition. Therefore this relation may only be used with reservation, and with the effect of BUE upon the real rake angle included. It can be seen that an increase in α will produce a decrease in γ. To determine the effect of BUE upon γ, a plastic-flow analysis according to Lee and

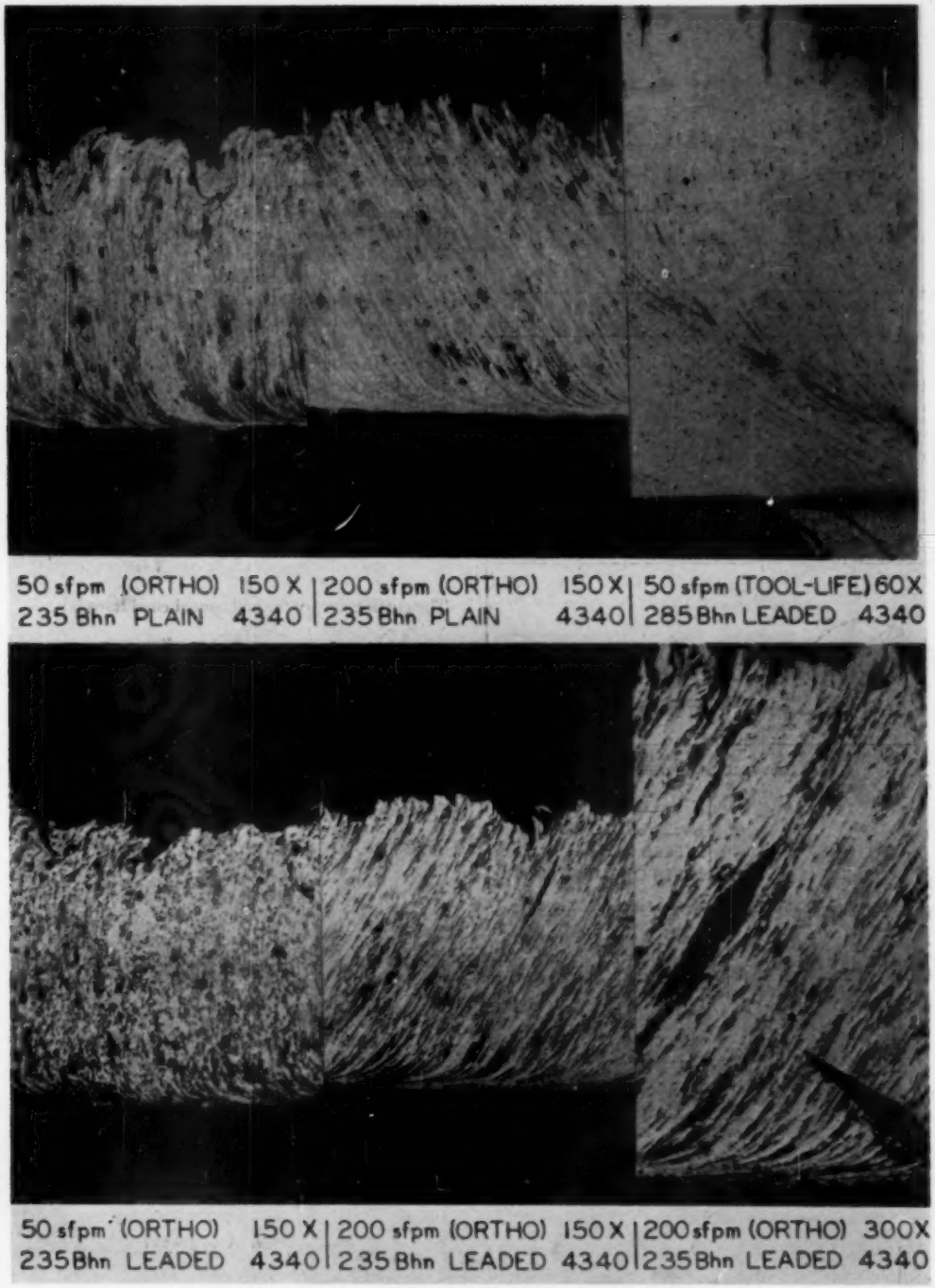

50 sfpm (ORTHO) 150 X
235 Bhn PLAIN 4340
200 sfpm (ORTHO) 150 X
235 Bhn PLAIN 4340
50 sfpm (TOOL-LIFE) 60X
285 Bhn LEADED 4340
50 sfpm (ORTHO) 150 X
235 Bhn LEADED 4340
200 sfpm (ORTHO) 150 X
235 Bhn LEADED 4340
200 sfpm (ORTHO) 300X
235 Bhn LEADED 4340

Fig. 8 Chip photomicrographs

Shaffer [7] was conducted. It was determined that all the shear (neglecting wipe-down) occurred in the region of the BUE and shear plane. From this it would be reasonable to obtain an average α and apply it to the shear-strain equation. The results of such an approximation using values obtained from Fig. 11 reveals that the shear strain of respective chips are

At 100 sfpm, 0.005 feed

$$\gamma_{plain} = \frac{\cos 5^\circ}{\sin 18^\circ \cos 13^\circ} = 3.31$$

$$\gamma_{leaded} = \frac{\cos 9^\circ}{\sin 28^\circ \cos 19^\circ} = 2.23$$

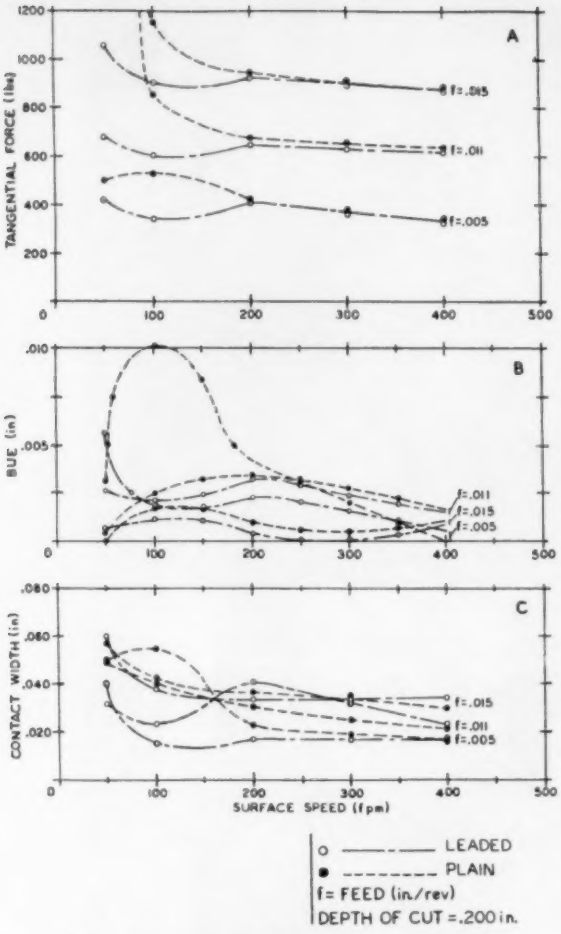


Fig. 9 Orthogonal BUE data at constant feeds (4340 steel)

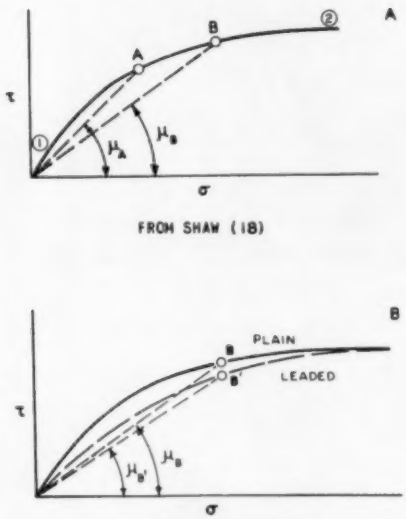


Fig. 10 Real versus apparent area theory

Thus the difference in chip hardness can be explained by the lead affecting the tool forces and shear angle.

Reviewing the foregoing discussion it appears that:

1 The lead smears on the tool face and provides lubrication at the tool-chip interface.

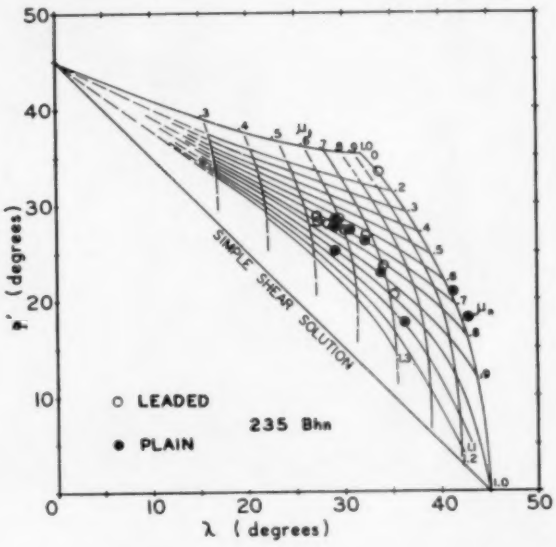
2 It was not determined whether or not the lead inclusions increased the number of active shear planes thereby producing a more homogeneous chip.

The critical speed displayed by leaded steel is important when it is to be considered for some specific machining operation. Whether the economic cutting speed lies above or below the critical speed is of prime importance in the final analysis since there is well-documented evidence [4, 12, 13, 14, 15, 21, 22] that the region characterized by decreased forces and μ is also one of increased tool life. Therefore it is advantageous to be able to determine the approximate location of the critical point as a function of the machining variables. The previous discussions of the critical point determined it to be a strong function of cutting speed and material strength. It was also found to be possibly a function of feed. This information suggests that the tool temperature may control the critical point. The commonly used relation for tool temperature [20] is

$$\theta_T = mU\sqrt{\frac{Vf}{k\rho c}}$$

The cutting energy for orthogonal conditions is

$$U = \frac{F_t}{fd}$$



THE FOLLOWING PARAMETERS WERE USED TO PLOT DATA (MERCHANT[10])

$$\tan\varphi' = \frac{\cos\alpha}{\frac{1}{r_c} - \sin\alpha} = r_c \qquad \tan\lambda = \frac{F_n + (F_t \tan\alpha)}{F_t - (F_n \tan\alpha)} = \frac{F_t}{F_t}$$

Fig. 11 Lee and Shaffer plasticity solution for α = 0 deg

$\theta_T \equiv$ tool temperature, deg C
$m \equiv$ constant
$U \equiv$ total cutting energy, in-lb per in.
$V \equiv$ cutting speed, fpm
$f \equiv$ feed, in. per rev
$k \equiv$ thermal conductivity of workpiece, cal sec per cm deg C
$\rho \equiv$ density of workpiece, gram per cu cm
$c \equiv$ specific heat of workpiece, cal per gram deg C
$d \equiv$ depth of cut, in.

Combining in terms of $\frac{\theta_T}{m}$ since m is a "constant"

$$\frac{\theta_T}{m} = F_s \sqrt{\frac{V}{f}} \left[\frac{0.069}{d\sqrt{k\rho c}} \right]$$

(See Nomenclature for units.)

From the experimental data the critical point $\left[\left(\frac{\theta_T}{m}\right)_{crit}\right]$ was found to have a value of $\simeq 8 \times 10^4$.

4340 Steel		0.20 depth
Bhn	Feed	$\left(\frac{\theta_T}{m}\right)_{crit}$
235	0.005	9.9×10^4
235	0.015	11.8×10^4
444	0.005	6.8×10^4

Boston [2] gives a relation for the tangential force (F_s) which holds approximately true at and above the critical speed,

$$F_s = C_0 f^a d^b$$

and permits a solution of the temperature parameter in terms of feed.

$$\frac{\theta_T}{m} = 0.069\, C_0 f^{a-1/2} d^{b-1} \sqrt{\frac{V}{k\rho c}}$$

where C_0 = function of strength level.

Substituting average values for the conditions of this report

$$a = 0.8 \qquad b = 1.0$$

Therefore

$$\frac{\theta_T}{m} = 0.069 C_0 f^{0.3} \sqrt{\frac{V}{k\rho c}}$$

This relationship is approximately true if the machining process does not deviate greatly from orthogonal cutting.

$$V_{crit} = 1.1 \times 10^3 \left[\frac{\sqrt{k\rho c}}{C_0 f^{0.3}} \right]^2$$

Solving for V_{crit}, it can be seen that the critical speed varies with the square of strength and less than proportionally to a change in feed.

Having already seen that applications with low economic cutting speeds tend to be conducive to the use of leaded steel, we wish also to determine the effects of feed and hardness. If the Taylor tool-life equation ($VT^n f^a = \xi$) holds in the region of interest, the economic tool life is given by

$$T = \Re\left(\frac{1}{n} - 1\right) \text{ where } \Re = \frac{xT_d + y}{x} \text{ and}$$

x = cost of machine and operator, cents per min
T_d = time to change tool, min
y = cost of tool per cutting edge, cents

Solving for the economic cutting speed (V_{min})

$$V_{min} = \frac{\xi}{f^a} \left[\Re\left(\frac{1}{n} - 1\right) \right]^{-n}$$

where a representative value of a is 0.8. Comparing this relationship with the expression for V_{crit}, it can be concluded that an increase in feed decreases V_{min} more than V_{crit} and therefore a high-feed application is more likely to benefit from leading the steel. Because a high $\Re$ decreases the economic cutting speed, a low machine and overhead cost or a high tool cost tend to increase the probability that leaded steel is advantageous. It is difficult to make a general conclusion as to the effect of strength level on use of leaded steel because the constants of the tool-life equation vary with strength.

The above discussion may be a helpful method of determining when leaded steel can be applied to increase tool life. Because the cost of leaded steel is greater, it is not always true that higher tool life is sufficient to substantiate its use. An additional analysis such as given by Cook [5] is required for this purpose.

Conclusions

1 The effects of lead on the cutting mechanism were revealed only below a critical speed, which was found to decrease for increased workpiece hardness.

2 The effects of lead include decreased machining coefficient of friction, increased shear angle, decreased shear energy, and lower cutting forces below the critical speed.

3 An increase in cutting speed tended to decrease tool life more when machining leaded than when machining plain steel.

4 The lead additive acting as a lubricant on the chip-tool interface was found to be the mechanism whereby the above effects were produced below the critical speed.

5 It is proposed that the critical speed is directly related to the tool temperature.

References

1 F. L. Bagley, Jr., and R. Mennell, "Machining Characteristics of Leaded Steel," Rodman Laboratory Report no. 23/6, Watertown Arsenal, February, 1959, p. B-1.

2 O. W. Boston, "Metal Processing," John Wiley & Sons, Inc., New York, N. Y., 1941, pp. 118–121.

3 F. W. Boulger and others, "Final Report on An Evaluation of the Present Understanding of Metal Cutting," Battelle Memorial Institute, Columbus, Ohio, August 30, 1958, p. 226.

4 F. W. Boulger and H. J. Grover, "Machinability Can Be Related to Composition," *Tool Engineer*, vol. 40, March, 1958, pp. 114–115.

5 N. H. Cook, "Machining Economics," M.I.T.-ASTE Regional Conference, May 9, 1959, published by Mechanical Engineering Department, M.I.T.

6 I. Finnie and M. C. Shaw, "The Friction Process in Metal Cutting," TRANS. ASME, vol. 78, 1956, pp. 1649–1657.

7 E. H. Lee and B. W. Shaffer, "The Theory of Plasticity Applied to a Problem of Machining," *Journal of Applied Mechanics*, vol. 18, TRANS. ASME, vol. 73, 1951, pp. 405–413.

8 R. F. Mennell and E. A. Jeffery, "Evaluation of Several Lathe Tool Life Testing Techniques," ASTE Semi-Annual Conference, Paper No. 216, vol. 59, book 2, October, 1959.

9 R. Mennell, "Comparison Tool Test Saves Time," *Tool Engineer*, vol. 43, September, 1959, pp. 121–124.

10 M. E. Merchant, "Mechanics of the Metal-Cutting Process, Part I—Orthogonal Cutting and a Type-2 Chip; Part II—Plasticity Conditions in Orthogonal Cutting," *Journal of Applied Physics*, vol. 16, May, 1945, pp. 267–275; June, 1945, pp. 318–324.

11 M. J. Moroney, "Facts From Figures," Penguin Books Inc., England, 1956, chapters 13 and 16.

12 W. J. Phillips and D. B. Barron, "Lead in Cast Steel, How Does Machinability Improve?" *Iron Age*, vol. 180, August 15, 1957, pp. 102–171.

13 W. I. Pumphrey, "Metallurgical Aspects of Machining of Steel," *Metal Progress*, vol. 64, August, 1953, pp. 143–144.

14 F. J. Robbins, "Four Factors Determine Selection of Leaded Steel," *Iron Age*, vol. 174, October 7, 1954, pp. 117–120.

15 M. C. Shaw and others, "The Influence of Lead on Metal-

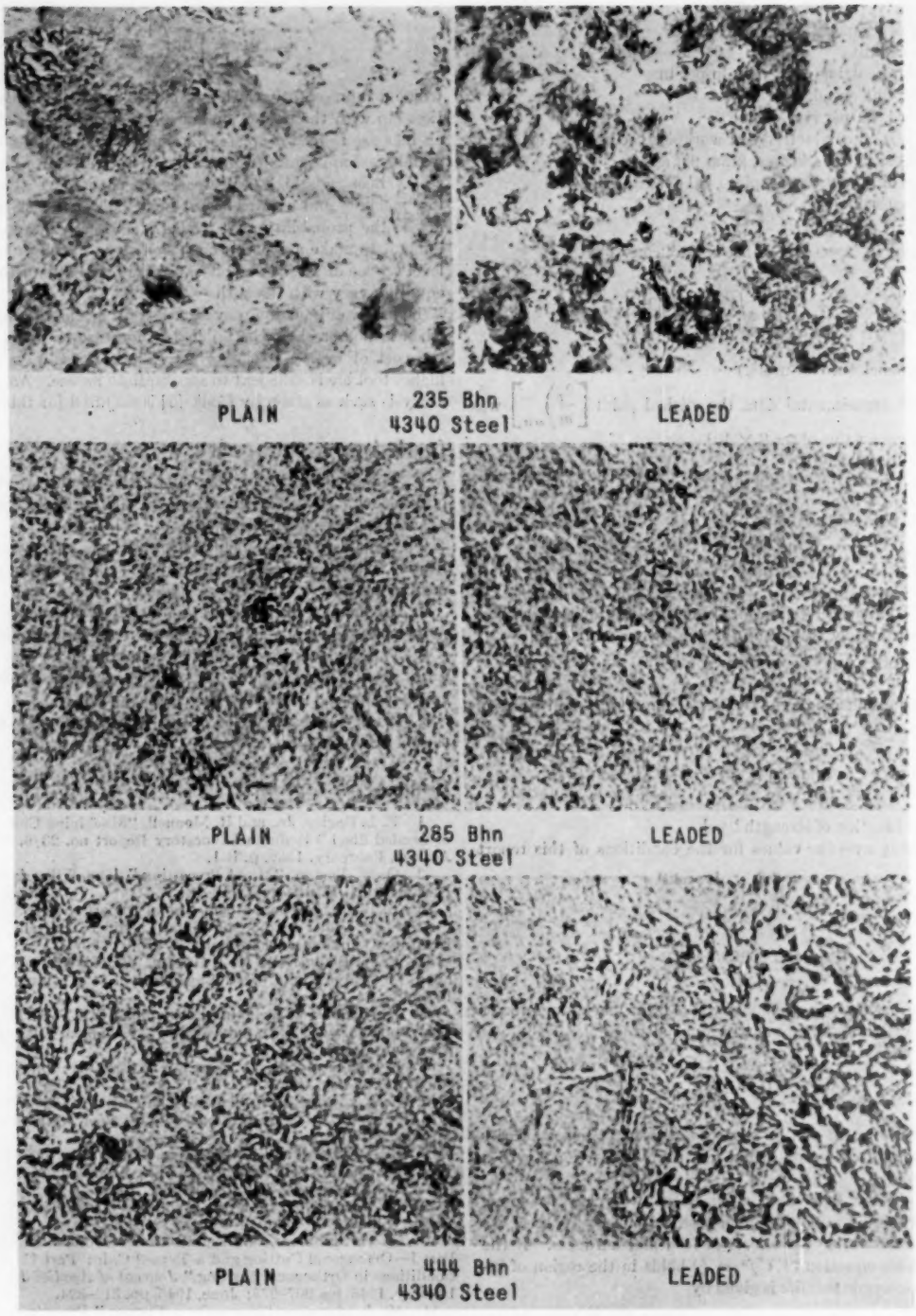


Fig. 12 Longitudinal stock photomicrographs (1000 X)

Cutting Forces and Temperature," TRANS. ASME, vol. 75, 1957, pp. 1143–1154.

16 M. C. Shaw and others, "Leaded Steel and the Real Area of Contact in Metal-Cutting," TRANS. ASME, vol. 79, 1957, pp. 1165–1171.

17 M. C. Shaw, "Metal-Cutting Principles," third edition, Massachusetts Institute of Technology, Cambridge, Mass., Chapter 13.

18 M. C. Shaw, Discussion of B. T. Chao and K. J. Trigger, "Controlled-Contact Cutting Tools," TRANS. ASME, Series B, *Journal of Engineering for Industry*, vol. 81, 1959, pp. 149–150.

19 M. C. Shaw, "On the Action of Metal Cutting Fluids at Low Speeds," *Wear*, vol. 2, Elsevier Publishing Company, Amsterdam, N. Y.

20 "Wear Theory," M. I. T. Metal Processing Division.

21 E. J. Weller, "Machinability of Type A Leaded Steels," *Metal Progress*, vol. 72, November, 1957, p. 196.

22 N. Zlatin and V. J. Gould, "The Effect of a Lead Additive on the Machinability of Alloy Steels," TRANS. ASME, Series B, *Journal of Engineering for Industry*, vol. 81, 1959, pp. 131–138.

APPENDIX

Physical Properties

The physical properties of a steel have pronounced effects upon its machining properties, therefore, this appendix was compiled primarily to facilitate correlation of the data and results presented in this report with that of other investigators.

The steel used was a split heat of 4340 hot-rolled 6-in. bar stock with a chemical composition as follows:

Element	Percentage
Carbon	0.39
Manganese	0.74
Phosphorus	0.012
Sulfur	0.022
Silicon	0.28
Nickel	1.73
Chromium	0.83
Molybdenum	0.22
Lead content of lead portion	0.15

The heat-treatments used are described in the following table:

Table of heat-treatment

Treatment	Time, hr	Temp, deg F	Water quench	Oil quench
235 Bhn				
Anneal	5	1500	Furnace cooled	
285 Bhn				
Austenitizing	5	1500	1500 to 500 F	500 F to room temp
Temper	6	1100	1100 F to room temp	
444 Bhn				
Austenitizing	5	1500	1500 to 500 F	500 F to room temp
Temper	6	700	700 F to room temp	

The longitudinal microstructure of each of these hardness levels is given in Fig. 12. The photomicrographs were taken from the midradius of the 6-in., hot-rolled, bar stock at a distance of 6 in. from the end of the heat-treated bar.

The 235-Bhn material consisted primarily of ferrite (white areas) and pearlite (dark areas), while the 285 and 444-Bhn specimens consisted of tempered martensite (dark areas), tempered bainite (gray areas), and ferrite (white areas). The banding consists of alternate concentrations of ferrite and was produced by the hot-rolling process. These bands will have no adverse effect upon the machining process since the cutting was done across them.

A General Flow Graph Technique for the Solution of Multiloop Sampled Systems

R. ASH
W. H. KIM
G. M. KRANC
Department of Electrical Engineering, Columbia University, New York, N. Y.

This paper presents a flow graph approach for analyzing multiloop sampled-data systems. Two techniques for finding the sampled output are examined. These are, first, construction of a "sampled" signal flow graph from the original system; second, a general gain formula which can be directly applied to the original system. The theorems necessary to establish the formula are rigorously proved.

The first technique allows two possible modes of solution. The sampled output can be found directly from the sampled flow graph by the use of Mason's formula, or in case of a more complicated multiloop system, the problem of enumerating nontouching feedback loops can be simplified by the use of topological matrices. Techniques developed in the paper are also applied to the solution of multirate systems.

Introduction

A NUMBER OF AUTHORS have used the signal flow graph [1][1] in analyzing a multiloop sampled-data system [2, 3, 4]. Recently, Lendaris and Jury [5] have found a general gain expression which is an extension of Mason's result for a linear continuous system. However, it turns out that the formula of Lendaris and Jury is not valid in general. This paper describes a general signal flow graph technique for analyzing multiloop sampled-data systems. Two techniques for finding the sampled output are developed. These are, first, the construction of a sampled signal flow graph from the original system. The sampled output may then be found from the sampled flow graph by the direct application of any of the known techniques for the solution of multiloop continuous systems. In particular, the authors suggest the use of Mason's general gain formula [1] or "topological" matrices [6]. The second technique developed in this paper allows the computation of the sampled output directly from the original system through the use of a general gain formula. Techniques for finding the unsampled output are also discussed.

To formulate the problem let us examine the system whose block diagram is shown in Fig. 1. In this system a plant is controlled by means of a bypass digital controller [7, 8], the whole structure forming a single sampled-loop system.

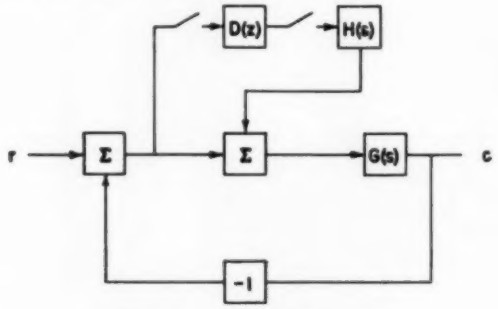


Fig. 1 Single sampled-loop system

A flow graph representation of the system of Fig. 1 is shown in Fig. 2 where for simplicity $D(z)$ is denoted by D^* and $G(s)$ and $H(s)$ by G and H, respectively. x_1, x_2, and x_3 are transforms of node variables appearing at nodes 1, 2, and 3. Nodes are placed at the intersection of branches and before and after a sampler in each sampled branch.[2] (If there are two or more samplers in a branch, they can always be reduced to one sampler followed by a sampled transfer function.)

The system of Fig. 2 is characterized by the equations

$$\left.\begin{aligned} x_1 &= r - x_3 \\ x_2 &= x_1^* \\ x_3 &= D^*GHx_2 + Gx_1 \end{aligned}\right\} \qquad (1)$$

Finding the output of the system in terms of the input and the system parameters is equivalent to solving equations (1) for x_3 or x_3^*. We shall use the sampled signal flow graph to simplify the solution.

Derivation of the Sampled Graph

In general, using the principle of superposition, the equations describing a multisampled-loop system may be written in matrix form as

$$AX = BX^* + R \qquad (2)$$

where A and B are square matrices whose order is equal to the number of variables, $\mathbf{X}$ is the column vector of variables, and $\mathbf{X}^*$ is obtained from $\mathbf{X}$ by replacing each variable by its sampled form. If Q is any matrix, the symbol Q^* will be used to denote the matrix obtained from Q by replacing each element q_{ij} of Q

[2] We shall use distinct symbols for the input and output variables of a sampler. If x_i is the input and x_j the output, then $x_j = x_i^*$. This procedure simplifies proofs which follow.

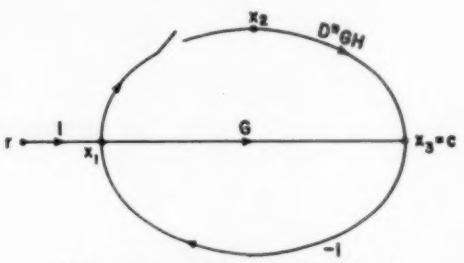


Fig. 2 Flow graph for the system in Fig. 1

[1] Numbers in brackets designate References at end of paper.

Contributed by the Instruments and Regulators Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASME-AIEE-IRE-ISA National Automatic Control Conference, Dallas, Texas, November 4–6, 1959.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 10, 1959. Paper No. 59—NAC-1.

by its sampled form q_{ij}^*. If we restrict ourselves to single-input systems (multiple input systems may be easily handled by superposition) and adopt the convention that the input is connected to the node corresponding to the variable x_1, then

$$\mathbf{R} = \begin{bmatrix} r \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

where r is the input. For the system of Fig. 1

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -G & -D^*GH & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \mathbf{X}^* = \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix}$$

If we solve equation (2) for X, we obtain

$$\mathbf{X} = A^{-1}B\mathbf{X}^* + A^{-1}\mathbf{R} \tag{3}$$

Taking the sampled form of equation (3) we have

$$\mathbf{X}^* = (A^{-1}B)^*\mathbf{X}^* + (A^{-1}\mathbf{R})^* \tag{4}$$

Equation (4) represents a set of simultaneous linear equations involving only the sampled variables as unknowns. The signal flow graph corresponding to equation (4) will be called the *sampled graph*. To construct the sampled graph we need to know the elements of the matrices $A^{-1}B$ and $A^{-1}\mathbf{R}$. The physical interpretation of A^{-1} is described in the next section. For the system of Fig. 1 we have

$$\left.\begin{aligned} A^{-1} &= \frac{1}{1+G}\begin{bmatrix} 1 & -D^*GH & -1 \\ 0 & 1+G & 0 \\ G & D^*GH & 1 \end{bmatrix} \\ A^{-1}B &= \frac{1}{1+G}\begin{bmatrix} -D^*GH & 0 & 0 \\ 1+G & 0 & 0 \\ D^*GH & 0 & 0 \end{bmatrix} \\ A^{-1}\mathbf{R} &= \frac{1}{1+G}\begin{bmatrix} r \\ 0 \\ Gr \end{bmatrix} \end{aligned}\right\} \tag{5}$$

Therefore, equation (4) can be written as

$$\left.\begin{aligned} x_1^* &= -D^*\left(\frac{GH}{1+G}\right)^* x_1^* + \left(\frac{r}{1+G}\right)^* \\ x_2^* &= x_1^* \\ x_3^* &= D^*\left(\frac{GH}{1+G}\right)^* x_1^* + \left(\frac{Gr}{1+G}\right)^* \end{aligned}\right\} \tag{6}$$

The sampled flow graph for this system is shown in Fig. 3. The output c^* computed from the sampled flow graph is

$$c^* = \left(\frac{r}{1+G}\right)^* \frac{D^*\left(\frac{GH}{1+G}\right)^*}{1+D^*\left(\frac{GH}{1+G}\right)^*} + \left(\frac{Gr}{1+G}\right)^* \tag{7}$$

If $A = G/(1+G)$, then equation (7) becomes upon simplification

$$c^* = \frac{(Ar)^* + (AH)^*D^*r^*}{1 + (AH)^*D^*} \tag{8}$$

which agrees with the result for the same problem obtained on page 188 of reference [8].

Efficient Construction of the Sampled Graph

In this section we shall demonstrate that it is always possible to obtain the sampled graph without the necessity of matrix inversion. The required elements of $A^{-1}B$ and $A^{-1}\mathbf{R}$ may be found by conventional signal-flow graph technique, as the following argument shows.

Let $A = [a_{ij}]$. The matrix A then characterizes the equation

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= y_2 \\ \cdot \\ \cdot \\ \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= y_n \end{aligned}\right\} \tag{9}$$

We may interpret these equations as describing the original control system with all samplers open-circuited. The forcing function y_k may be interpreted as a source connected to node k of the original graph. If a_{ij}^{-1} is the element in row i and column j of the matrix A^{-1}, then

$$a_{ij}^{-1} = \frac{A_{ji}}{|A|} = \left.\frac{x_i}{y_j}\right]_{\text{all other } y\text{'s} = 0} \tag{10}$$

where

A_{ji} = the cofactor of a_{ji}

$|A|$ = the determinant of A

Thus a_{ij}^{-1} is the response at node i of the original system due to a unit source applied at node j, with all samplers open circuited. Let

$$a_{ij}^{-1} = G_{ji} \tag{11}$$

where G_{ji} is the transfer function from node j to node i of the original system with all samplers open-circuited.

The following properties of the matrix B are easily verified.

1 The elements of B are either 1 or 0.

2 The element b_{ij} of B will be 1 if and only if the variable x_j is the input to a sampler and x_i is the output of the same sampler.

3 B contains at most one nonzero element per column. (Two samplers with the same input may always be replaced by a single sampler.)

4 If $b_{ij} = 1$, then column i of A^{-1} appears as column j of $(A^{-1}B)$.

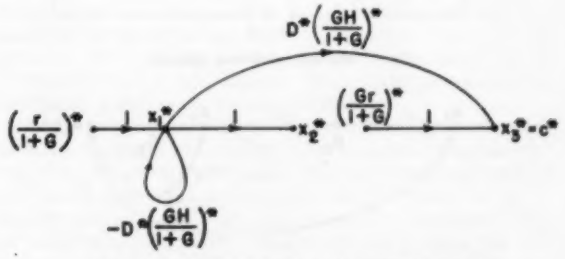


Fig. 3 Sampled flow graph for the system in Fig. 1

5 If x_j is not the input to a sampler, column j of $(A^{-1}B)$ consists entirely of zeros.

From these properties and equation (11) we may make the following statement:

If the variable x_j is the input to a sampler and the variable x_i is the output of the same sampler, then column j of $A^{-1}B$ appears as follows:

$$A^{-1}B = \begin{bmatrix} --- G_{i1} --- \\ --- G_{i2} --- \\ --- G_{in} --- \end{bmatrix} \tag{12}$$

If x_j is not the input to a sampler, then column j of $A^{-1}B$ consists entirely of zeros.

From equation (11), the matrix $A^{-1}\mathbf{R}$ has the form

$$A^{-1}\mathbf{R} = \begin{bmatrix} G_{11}r \\ G_{12}r \\ \cdot \\ \cdot \\ \cdot \\ G_{1n}r \end{bmatrix} \tag{13}$$

Equations (12) and (13) allow the sampled graph to be constructed efficiently from the original graph.

Illustrative Example I

To illustrate the technique of efficient sampled flow graph construction let us consider a multisampled-loop system shown in Fig. 4. The flow graph representation of this system is shown in Fig. 5.

We observe that x_1 and x_6 are the only variables at the input to samplers. Therefore we can write directly from Fig. 5 that

$$A^{-1}B = \begin{bmatrix} G_{21} & 0 & 0 & 0 & 0 & G_{61} \\ G_{22} & 0 & 0 & 0 & 0 & G_{62} \\ G_{23} & 0 & 0 & 0 & 0 & G_{63} \\ G_{24} & 0 & 0 & 0 & 0 & G_{64} \\ G_{25} & 0 & 0 & 0 & 0 & G_{65} \\ G_{26} & 0 & 0 & 0 & 0 & G_{66} \end{bmatrix} \tag{14}$$

Evaluating the elements of $A^{-1}B$ we have

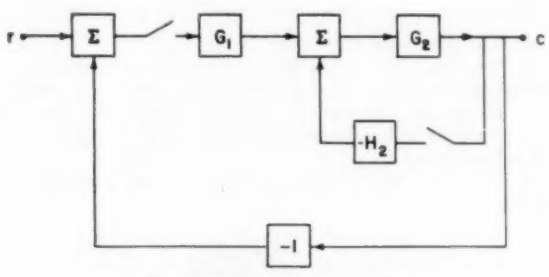


Fig. 4 Multisampled-loop system

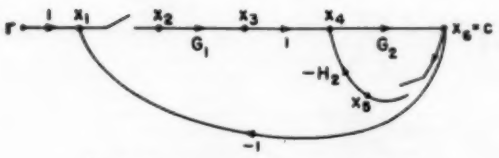


Fig. 5 Flow graph for the system in Fig. 4

$$A^{-1}B = \begin{bmatrix} -G_1G_2 & 0 & 0 & 0 & 0 & G_2H_2 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ G_1 & 0 & 0 & 0 & 0 & 0 \\ G_1 & 0 & 0 & 0 & 0 & -H_2 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ G_1G_2 & 0 & 0 & 0 & 0 & -G_2H_2 \end{bmatrix} \tag{15}$$

For the system under consideration, $G_{1i} = 0$ for $i \neq 1$ and therefore

$$A^{-1}\mathbf{R} = \begin{bmatrix} r \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{16}$$

In view of these relations, equation (4) can be represented by a sampled flow graph shown in Fig. 6. The output c^*, which can now be computed by the direct application of Mason's formula [1] or by the use of topological matrices [6], is

$$c^* = \frac{r^*(G_1G_2)^*}{1 + (G_1G_2)^* + (G_2H_2)^*} \tag{17}$$

Illustrative Example II (Multirate System)

In this example we shall illustrate the application of the sampled flow graph to the solution of systems containing samplers with periods T/p_1, T/p_2, T/p_3 . . . where p_i is a positive integer. It has been shown elsewhere [9] that each sampler operating with a period T/p_i can be replaced, using the "switch decomposition technique" by a system containing advance and delay elements together with a number of samplers all operating with period T. Thus any multirate system can be replaced by an equivalent single rate system possessing a more complicated structure. It has been shown [9, 10] that the transform technique of solution of such systems leads to simultaneous algebraic equations one of whose unknowns is the transform of the output. The use of the sampled flow graph avoids the problem of solution of simultaneous equations.

Without any loss of generality let us consider the simple multirate system shown in Fig. 7. The equivalent single-rate system is

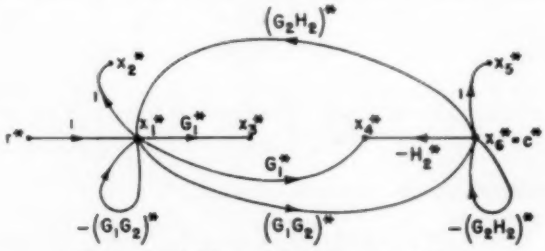


Fig. 6 Sampled flow graph for the system in Fig. 4

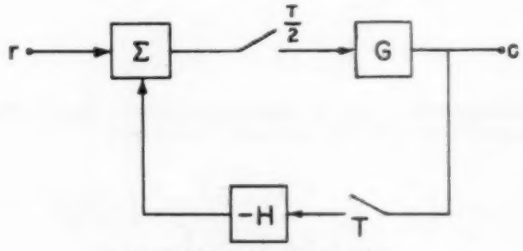


Fig. 7 A simple multirate system

shown in Fig. 8. Fig. 9 shows the flow graph representation of this system.

Since x_1, x_3, and x_4 are the only variables at the input to samplers, therefore

$$A^{-1}B = \begin{bmatrix} G_{11} & 0 & G_{61} & G_{11} & 0 & 0 \\ G_{22} & 0 & G_{62} & G_{62} & 0 & 0 \\ G_{31} & 0 & G_{61} & G_{61} & 0 & 0 \\ G_{34} & 0 & G_{64} & G_{64} & 0 & 0 \\ G_{35} & 0 & G_{65} & G_{65} & 0 & 0 \\ G_{36} & 0 & G_{66} & G_{66} & 0 & 0 \end{bmatrix} \tag{18}$$

Computing all G_{ij}'s from the graph we obtain

$$A^{-1}B = \begin{bmatrix} 0 & 0 & -H & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0_2G & 0 & 0 \\ 0 & 0 & -H0_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{19}$$

and similarly

$$A^{-1}\mathbf{R} = \begin{bmatrix} r \\ 0 \\ 0 \\ 0_1r \\ 0 \\ 0 \end{bmatrix} \tag{20}$$

Substituting the expressions for $A^{-1}B$ and $A^{-1}\mathbf{R}$ into equation (4) we obtain

$$\left.\begin{aligned} x_1^* &= -H^*x_3^* + r^* \\ x_2^* &= x_1^* \\ x_3^* &= G^*x_1^* + (0_2G)^*x_4^* \\ x_4^* &= -(H0_1)^*x_3^* + (0_1r)^* \\ x_5^* &= x_4^* \\ x_6^* &= x_3^* \end{aligned}\right\} \tag{21}$$

The sampled flow graph representing equations (21) is shown in Fig. 10. The expression for the sampled output, c^*, which can be computed using Mason's formula or topological matrices, is

$$c^* = \frac{r^*G^* + (0_1r)^*(0_2G)^*}{1 + G^*H^* + (0_2G)^*(0_1H)^*} \tag{22}$$

A General Gain Formula

It is possible to bypass the sampled graph entirely and compute the sampled form of the output directly from the original system. Rearranging equation (4) for convenience,

$$[I - (A^{-1}B)^*]X^* = (A^{-1}R)^* \tag{23}$$

where I is a unit matrix whose order is equal to that of A and B.

If the output c is the node variable x_n, then using equation (13), we have

$$c^* = x_n^* = \frac{(G_{11}r)^*\Delta_{1n}}{\Delta} + \frac{(G_{12}r)^*\Delta_{2n}}{\Delta} + \ldots + \frac{(G_{1n}r)^*\Delta_{nn}}{\Delta} \tag{24}$$

where

Δ = determinant of $I - (A^{-1}B)^*$

Δ_{in} = cofactor of the element in row i and column n of $I - (A^{-1}B)^*$

The determinant Δ and the various cofactors Δ_{in} may be computed by applying Mason's formula [1] to the sampled graph.

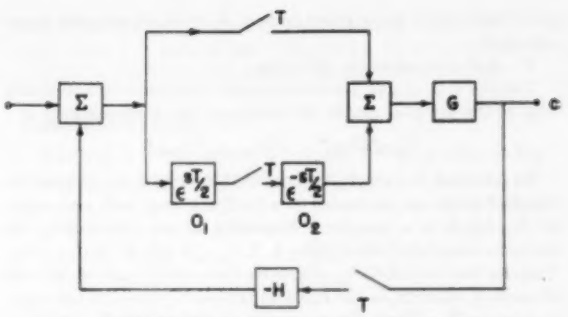


Fig. 8 Single-rate equivalent of the system in Fig. 7

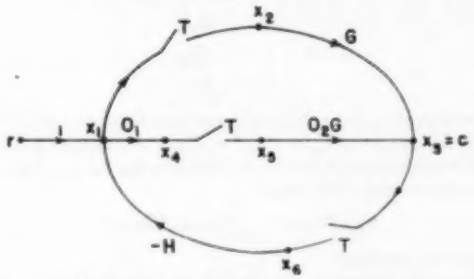


Fig. 9 Flow graph for the system in Fig. 8

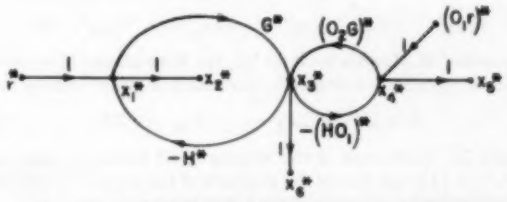


Fig. 10 Sampled flow graph for the system in Fig. 8

In order to avoid the construction of the sampled graph, we must know the mechanism by which the feedback loops and forward paths occurring in Mason's formula are formed.

(a) Feedback Loops. We shall first establish the following result:

If a branch originates from node x_j^* of the sampled graph, then x_j must be the input to a sampler in the graph representing the original system.

To justify this statement, suppose that a branch b_{jk} originates from node j and terminates on node k of the sampled graph. From (4) and (13) the equation at node k of the sampled graph is

$$x_k^* = t_{1k}^*x_1^* + t_{2k}^*x_2^* + \ldots + t_{jk}^*x_j^* + \ldots + t_{nk}^*x_n^* + (G_{1k}r)^* \tag{25}$$

where t_{jk}^* is the element in row k and column j of $(A^{-1}B)^*$ and is also the weight of the branch b_{jk}. From the previously discussed properties of the matrix B, it follows that if x_j is not the input to a sampler, then column j of $(A^{-1}B)$ consists entirely of zeros, and hence $t_{jk}^* = 0$. Thus if x_j is not the input to a sampler, no branch b_{jk} can exist in the sampled graph.

We shall now adopt the convention of numbering the samplers so that x_j is the input to sampler j. For example, if a system contains three samplers, whose inputs are x_1, x_3, and x_6, then we shall call the samplers S_1, S_3, and S_6, respectively. We shall denote by G_{x_jk} the gain from the *output* node of sampler S_j to node k, with all samplers open-circuited. We shall denote by $G_{s_is_j}$ the gain from the *output* node of S_i to the *input* node of S_j, and G_{is_q} the

gain from node i to the *input* node of S_q, with all samplers open-circuited.

We shall now show the following:

If nodes 1, 2, . . ., m, 1 are connected *in that order by* a feedback loop of the sampled graph, the weight of the feedback loop is

$$G_{s_1s_2}{}^*G_{s_2s_3}{}^* \ldots G_{s_{m-1}s_m}{}^*G_{s_ms_1}{}^*$$

To establish this result, we note that if nodes 1, 2, . . . m of the sampled graph are connected by a feedback loop, each node must be the input to a sampler. According to our convention, the samplers associated with nodes 1, 2, . . ., m are S_1, S_2, . . ., S_m. Suppose that a branch b_{jk} originates from node j and terminates on node k, where x_j is the input to sampler S_j and k is the input to sampler S_k. From the properties of the matrix B, and from (12)$_j$ column j of $(A^{-1}B)$ appears as follows:

$$A^{-1}B = \begin{bmatrix} --- G_{s_j1} --- \\ --- G_{s_j2} --- \\ --- G_{s_jn} --- \end{bmatrix} \tag{26}$$

The weight of b_{jk} is the element in row k and column j of $(A^{-1}B)^*$. Hence the weight of $b_{jk} = G_{s_jk}{}^* = G_{s_js_k}{}^*$.[1] The feedback loop under consideration is formed by branches b_{12}, b_{23}, . . ., $b_{m-1,m}b_{m1}$. Hence the weight of the loop is

$$G_{s_1s_2}{}^*G_{s_2s_3}{}^* \ldots G_{s_{m-1}s_m}{}^*G_{s_ms_1}{}^*,$$

as has been asserted.

From this discussion, the weight of a feedback loop of the sampled graph must be of the following form:

$$G_{s_is_i}{}^*, \quad G_{s_is_j}{}^*G_{s_js_i}{}^*, \quad G_{s_is_j}{}^*G_{s_js_k}{}^*G_{s_ks_i}{}^*, \quad \text{etc.}$$

According to Mason's formula [1], the determinant of a system of linear equations described by a signal flow graph is given by

$$1 - \Sigma P_1 + \Sigma P_2 - \ldots + (-1)^m \Sigma P_m \tag{27}$$

where ΣP_1 is the sum of the weights of all feedback loops, and ΣP_i $(i > 1)$ is the sum of the products of the weights of the nontouching feedback loops, taken i at a time.

Applying Mason's formula to the sampled graph, we obtain the following expression for Δ in equation (24):

$$\Delta = 1 - \left[\sum_i G_{s_is_i}{}^* + 1/2 \sum_{i,\,j} G_{s_is_j}{}^*G_{s_js_i}{}^* + 1/3 \sum_{\substack{i,\,j,\,k \\ i \neq j \\ i \neq k \\ j \neq k}} G_{s_is_j}{}^*G_{s_js_k}{}^*G_{s_ks_i}{}^* + \ldots \right] + \Sigma P_2 - \Sigma P_3 + \ldots + (-1)^m \Sigma P_m \tag{28}$$

where ΣP_k is the sum of the products of the weights of the *distinct nontouching* feedback loops in brackets, taken k at a time. The number m is the maximum number of distinct nontouching loops. Two loops of the sampled graph will be nontouching if the expressions for the weights of the loops, as appearing in (28), have no indexes in common. For example, the terms $G_{s_1s_2}{}^*G_{s_2s_1}{}^*$ and $G_{s_3s_4}{}^*G_{s_4s_3}{}^*$ correspond to nontouching loops.

The appearance of the factors 1/2, 1/3, etc., in (28) may be explained by noting that the terms

$$G_{s_1s_2}{}^*G_{s_2s_1}{}^*, \quad G_{s_2s_3}{}^*G_{s_3s_2}{}^*, \quad G_{s_3s_1}{}^*G_{s_1s_3}{}^*$$

all represent the same feedback loop. All three terms appear, however, in

$$\sum_{\substack{i,\,j,\,k \\ i \neq j \\ i \neq k \\ j \neq k}} G_{s_is_j}{}^*G_{s_js_k}{}^*G_{s_ks_i}{}^*$$

(b) Forward Paths. Let us now describe the nature of forward paths of the sampled graph. Suppose we wish to find all forward paths from node i to the output node (node n). For a forward path to exist at all node i must be an input to a sampler, which by convention we call S_i. Then the following terms may appear as weights of forward paths:

$$G_{s_in}{}^*, \quad G_{s_is_1}{}^*G_{s_1n}{}^*, \quad G_{s_is_2}{}^*G_{s_2n}{}^*, \quad G_{s_is_1}{}^*G_{s_1s_2}{}^*G_{s_2n}{}^*, \quad G_{s_is_2}{}^*G_{s_2s_1}{}^*G_{s_1n}{}^* \tag{29}$$

etc.

Note that the sampler S_n will never appear in (28), since the last branch of a forward path always terminates on node n. The only forward path from node n to node n has weight 1.

According to Mason's formula [1], the cofactor of the element in row i and column n of the determinant of a system of linear equations described by a signal flow graph is given by

$$\sum_j F_j\Delta_j \tag{30}$$

where F_j is the weight of a forward path from node i to node n, and Δ_j is the expression (27) evaluated for the graph obtained by removing all branches touching any of the nodes of F_j.

Again applying Mason's formula to the sampled graph, we obtain the following expression for Δ_{in} in (24).

$$\Delta_{in} = \sum_j F_j\Delta_j \tag{31}$$

where F_j is one of the terms in (29) and Δ_j is the expression (28) evaluated with all terms involving S_n or samplers appearing in F_j deleted.

The terms $(G_{1i}r)^*$ in equation (24) need be computed only if node i is the input to a sampler. Accordingly, for each sampler S_i, we shall have to compute a term $\dfrac{(G_{1s_i}r)^*\Delta_{in}}{\Delta}$ in (24).

Illustrative Example III

As an example, consider the system shown in Fig. 11. Following Lendaris and Jury [5], define

$$L_1 = \frac{G_1}{1 + G_1G_2G_3H_1} \qquad L_2 = \frac{G_1G_2}{1 + G_1G_2G_3H_1}$$

$$L_3 = \frac{G_1G_2G_3}{1 + G_1G_2G_3H_1}$$

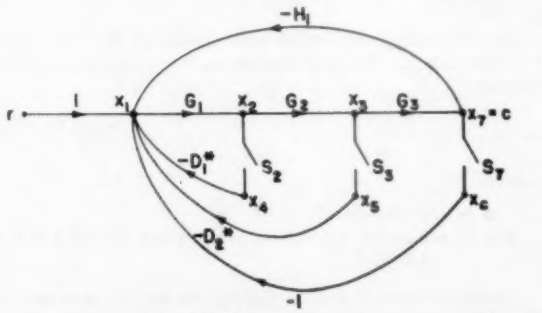


Fig. 11 Illustrative example III

[1] In view of (25), and the fact that $t_{jk}{}^* = 0$ if x_j is not the input to a sampler and $t_{jk}{}^* = G_{s_jk}{}^*$ if x_j is the input to sampler S_j, the sampled graph may be constructed directly from the original system without writing any equations. The procedure previously described for the construction of the sampled graph seems to be more systematic and therefore preferable in more complicated systems.

By (28)

$$\Delta =$$

$$1 - [G_{s_3s_4}^* + G_{s_2s_4}^* + G_{s_1s_7}^* + G_{s_2s_4}^* G_{s_3s_4}^* + G_{s_3s_7}^* G_{s_3s_4}^* + G_{s_2s_7}^* G_{s_3s_4}^* + G_{s_3s_4}^* G_{s_3s_4}^*] + G_{s_2s_4}^* G_{s_3s_4}^* + G_{s_2s_3}^* G_{s_3s_7}^* + G_{s_2s_4}^* G_{s_3s_7}^* + G_{s_2s_4}^* (G_{s_3s_7}^* G_{s_3s_4}^*) + G_{s_3s_4}^* (G_{s_2s_7}^* G_{s_3s_4}^*) + G_{s_2s_7}^* (G_{s_2s_4}^* G_{s_3s_4}^*) - G_{s_2s_4}^* G_{s_3s_4}^* G_{s_3s_7}^* \quad (32)$$

By inspection of the original system we obtain

$$\left.\begin{aligned} G_{s_1s_4}^* &= -D_1^* L_1^*, \quad G_{s_2s_4}^* = -D_2^* L_1^*, \quad G_{s_7s_4}^* = -L_1^* \\ G_{s_1s_3}^* &= -D_1^* L_2^*, \quad G_{s_2s_3}^* = -D_2^* L_2^*, \quad G_{s_7s_3}^* = -L_2^* \\ G_{s_1s_7}^* &= -D_1^* L_3^*, \quad G_{s_2s_7}^* = -D_2^* L_3^*, \quad G_{s_7s_7}^* = -L_3^* \end{aligned}\right\} \quad (33)$$

Substituting (33) into (32) we obtain

$$\Delta = 1 - [-D_1^* L_1^* - D_2^* L_2^* - L_3^* + D_1^* L_2^* D_2^* L_1^* + D_1^* L_3^* L_1^* + D_2^* L_3^* L_2^* - D_1^* L_2^* D_2^* L_3^* L_1^* - D_1^* L_3^* L_2^* D_2^* L_1^*] + D_1^* L_1^* D_2^* L_2^* + D_1^* L_1^* L_3^* + D_2^* L_2^* L_3^* - 3D_1^* D_2^* L_1^* L_2^* L_3^* + D_1^* D_2^* L_1^* L_2^* L_3^*$$

Simplifying,

$$\Delta = 1 + D_1^* L_1^* + D_2^* L_2^* + L_3^* \quad (34)$$

The numerator of c^* is, by (24),

$$(G_{1s_4} r)^* \Delta_{47} + (G_{1s_3} r)^* \Delta_{37} + (G_{1s_7} r)^* \Delta_{77} \quad (35)$$

We have $G_{1s_4} = L_1$, $G_{1s_3} = L_2$, $G_{1s_7} = L_3$.
By (31), (29), and (32),

$$\left.\begin{aligned} \Delta_{47} &= G_{s_4 7}^* (1 - G_{s_3 s_4}^*) + G_{s_3 s_4}^* G_{s_3 7}^* \\ \Delta_{37} &= G_{s_3 7}^* (1 - G_{s_3 s_3}^*) + G_{s_3 s_4}^* G_{s_4 7}^* \\ \Delta_{77} &= 1 - [G_{s_3 s_4}^* + G_{s_1 s_4}^* + G_{s_2 s_3}^* G_{s_3 s_4}^*] + G_{s_3 s_4}^* G_{s_3 s_4}^* \end{aligned}\right\} \quad (36)$$

From the original system

$$G_{s_4 7}^* = -D_1^* L_3^*, \qquad G_{s_3 7}^* = -D_2^* L_3^*$$

Hence (35) becomes

$$(rL_1)^* [-D_1^* L_3^* (1 + D_2^* L_2^*) + D_1^* L_2^* D_2^* L_3^*] + (rL_2)^* [-D_2^* L_3^* (1 + D_1^* L_1^*) + D_2^* L_1^* D_1^* L_3^*] + (rL_3)^* [1 - (-D_1^* L_1^* - D_2^* L_2^* + D_1^* L_2^* D_2^* L_1^*) + D_1^* L_1^* D_2^* L_2^*] = -(rL_1)^* D_1^* L_3^* - (rL_2)^* D_2^* L_3^* + (rL_3)^* (1 + D_1^* L_1^* + D_2^* L_2^*)$$

and finally

$$c^* = x_7^* = \frac{-(rL_1)^* D_1^* L_3^* - (rL_2)^* D_2^* L_3^* + (rL_3)^* (1 + D_1^* L_1^* + D_2^* L_2^*)}{1 + D_1^* L_1^* + D_2^* L_2^* + L_3^*}$$

$$= (rL_3)^* - \frac{(rL_1)^* D_1^* L_3^* + (rL_2)^* D_2^* L_3^* + (rL_3)^* L_3^*}{1 + D_1^* L_1^* + D_2^* L_2^* + L_3^*} \quad (37)$$

Computation of the Unsampled Output

The unsampled output may be obtained from the sampled graph using equation (3). In general it is necessary to compute *all* sampled node variables $x_1^*, x_2^*, \ldots, x_n^*$. The unsampled output is obtained by matrix multiplication using (3). It appears that, if a general gain formula for the unsampled output existed, it would be cumbersome and difficult to apply.

Illustrative Example IV

We shall find the unsampled output of the system of Fig. 1. From the sampled graph of Fig. 3 we obtain

$$\left.\begin{aligned} x_1^* &= x_2^* = \frac{\left(\frac{r}{1+G}\right)^*}{1 + D^* \left(\frac{GH}{1+G}\right)^*} \\ x_3^* &= \left(\frac{r}{1+G}\right)^* D^* \left(\frac{GH}{1+G}\right)^* \end{aligned}\right\} \quad (38)$$

From (3) and (5) we obtain

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{1+G} \begin{bmatrix} -D^*GH & 0 & 0 \\ 1+G & 0 & 0 \\ D^*GH & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix} + \frac{1}{1+G} \begin{bmatrix} r \\ 0 \\ Gr \end{bmatrix}$$

or

$$x_3 = c = \frac{D^*GH}{1+G} x_1^* + \frac{Gr}{1+G} \quad (39)$$

Substituting from (38) we obtain

$$x_3 = \frac{D^*GH}{1+G} \frac{\left(\frac{r}{1+G}\right)^*}{1 + D^* \left(\frac{GH}{1+G}\right)^*} + \frac{Gr}{1+G} \quad (40)$$

Conclusions

This paper has presented two systematic techniques which may be applied without restriction to any multiloop single or multirate sampled-data system. For complicated systems, the techniques are appreciably more efficient than the standard algebraic approach. In the opinion of the authors, it is easier to construct the sampled flow graph than to apply the general gain formula directly. It is found that in most cases the flow graph techniques are not "minimal"; expressions occur which contain identical terms with positive and negative signs. Such terms have no net contribution to the system output. This suggests the possibility of improving the efficiency of the techniques presented in this paper.

Acknowledgment

This research was supported in part by the U. S. Air Force under Contract No. AF 18(600)-677 monitored by the AF Office of Scientific Research of the Air Research and Development Command and in part by the National Science Foundation Grant G-6020.

References

1 S. J. Mason, "Feedback Theory—Further Properties of Signal Flow Graphs," *Proceedings, IRE*, vol. 44, 1956, pp. 920–926.

2 J. M. Salser, "Signal Flow Reductions in Sampled Data Sys-

tems," 1957 IRE Wescon Convention Record, Part 4, pp. 166–170.
3 W. K. Linvill and R. W. Sittler, "Design of Sampled-Data Systems by Extension of Conventional Techniques," MIT Report No. R-222, July, 1957.
4 C. W. Johnson, D. P. Lindorff, and C. G. A. Nordling, "Extension of Continuous Data System Design Techniques to Sampled-Data Control Systems," *Trans. AIEE*, vol. 73, Part II, 1954, pp. 147–153.
5 G. G. Lendaris and I. J. Jury, "Input-Output Relationships for Multi-Sampled-Loop Systems," AIEE, Trans. Paper No. 59-882, to be published.
6 R. Ash, "Topology and the Solution of Linear Systems," Tech. Report T-31/B, Department of Electrical Engineering, Columbia University, June, 1958.
7 K. K. Maitra and P. E. Sarachik, "Digital Compensation of Continuous Data Feedback Control Systems," *Trans. AIEE*, vol. 76, Pt. II, No. 24, 1956.
8 J. R. Ragazzini, and G. Franklin, "Sampled Data Control Systems," McGraw-Hill Book Company, Inc., New York, N. Y., 1958, p. 186.
9 G. M. Krane, "Input-Output Analysis of Multirate Feedback Systems," *Trans. IRE*, PGAC-3, November, 1957.
10 G. M. Krane, "Additional Techniques for Sampled-Data Problems," IRE Wescon Convention Record, 1957.

DISCUSSION

Benjamin C. Kuo[4]

The authors have presented a very unique and useful signal flow graph approach for analyzing multiloop sampled-data systems. Once the system node variables are properly designated, the matrices $A^{-1}B$ and $A^{-1}R$ can be obtained by inspection from the original flow graph. The sampled flow graph is constructed by using $A^{-1}B$, $A^{-1}R$, and Equation (4). The sampled output c^* may then be computed from the sampled flow graph by the use of Mason's gain formula.

The authors pointed out that in order to compute the unsampled output c, it is in general necessary to compute all the sampled node variables $x_1^*, x_2^*, \ldots . x_n^*$. The unsampled variables, including the unsampled output c^* are then obtained by matrix multiplication using Equation (3). However, since Equation (3) indicates that the unsampled variables are related to the sampled variables, it suggests that a composite signal flow graph may be constructed which is composed of the original

[4] Coordinated Science Laboratory & Electrical Engineering Department, University of Illinois, Urbana, Ill.

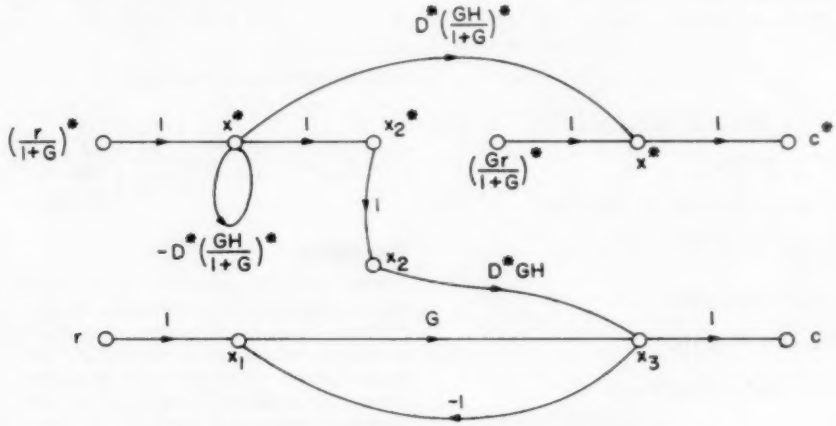


Fig. 12 Composite signal flow graph of sampled-data system in Fig. 1

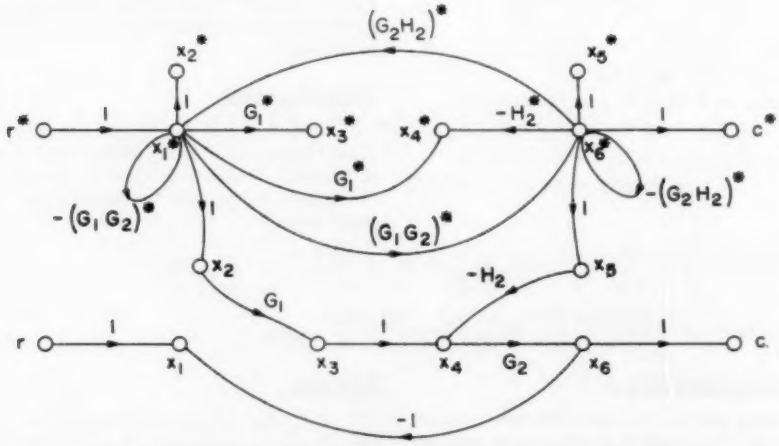


Fig. 13 Composite signal flow graph of sampled-data system in Fig. 4

flow graph and the sampled flow graph. For instance, the composite signal flow graph for the system in Fig. 1 is shown in Fig. 12. The composite graph consists of the sampled graph of Fig. 3 and the original graph of Fig. 2. The sampling switch of the original flow graph is omitted, and the sampling operation is described by drawing a branch of unity gain from node x_1^* to node x_2 representing the relation $x_2 = x_1^*$. From this composite flow graph, all the variables x_1, x_2, x_3, x_1^*, x_2^*, and x_3^* can be computed by means of Mason's formula. Thus from Fig. 12 the unsampled output is

$$x_3 = c = \frac{D^*GH\left(\frac{r}{1+G}\right)^*}{(1+G)\left[1+D^*\left(\frac{GH}{1+G}\right)^*\right]} + \frac{Gr}{1+G} \quad (41)$$

which agrees with Equation (40).

Similarly, the composite signal flow graph for the system in Fig. 4 is constructed in Fig. 13. The unsampled output is obtained directly by applying Mason's formula.

$$x_4 = c = \frac{G_1G_2[1+(G_2H_2)^*]r^* - r^*(G_1G_2)^*G_2H_2}{1+(G_1G_2)^*+(G_2H_2)^*} \quad (42)$$

It is apparent that all the other unsampled variables can also be obtained by the same procedure.

George G. Lendaris[5,6]

The authors have successfully formulated an accurate procedure for determining the *sampled* output of a general multi-sampled-loop system whose samplers are in synchronism (multiple or submultiple rates included). The discusser was quite pleased to note that another group of investigators also considered this an important problem.

The authors suggested in the conclusion of their paper that the efficiency of the techniques could perhaps be improved. It is the aim of this discussion to present an efficient technique for carrying out this procedure.

As the authors mentioned in their introduction, the discusser's work and resulting paper in this area[7] led to a formula that turned out not to be completely general. However, in pursuing this work further, the discusser developed an accurate general technique for determining the *sampled* and *continuous* output for these systems. This latter work was included in the paper[7] as a closure to supersede the method given in the main body of the paper.

In general, it is found that the topological method is more efficient than the algebraic formula method. This is true because most engineering problems are such that the system determinant contains some zero terms. The algebraic formula method by its nature does not take advantage of this fact before the calculations are made, whereas the topological method does. The latter does because for every zero term in the determinant, there is a branch missing in the flow graph; hence this term does not enter into the calculation procedure. For the case where all the terms of the determinant are nonzero, the algebraic and topological methods are equally as efficient (or inefficient, as the case may be).

The general approach to this problem (as the authors have also indicated) is to represent the given system block diagram by some sampled-signal flow graph, and then use standard techniques to proceed from the flow graph to the desired result. Assuming an efficient method for obtaining the desired result from the signal flow graph is available,[8] the problem reduces to that of finding an efficient signal flow graph representation of the given system.

In the discusser's opinion, for a system with n samplers, the best representation for the system should be a sampled signal flow graph with $n + 2$ nodes; one for each sampler, one for the input, and one for the output. The discusser derived such a form in the previously mentioned work,[7] and will present the results here.

The following is the proposed procedure for transforming the block diagram representation of the system into its efficient sampled signal flow graph representation.

(1) Eliminate all redundant samplers in the original block diagram. This step is not necessary for correct results, but it simplifies the resulting flow graph.

(2) Number all samplers (in any fashion desired).

(3) On a separate sheet of paper draw a node for each sampler indicated in step (2) and label these S_1^*, S_2^*, . . ., S_n^*.

(4) Open all samplers in the block diagram.

(5) Start from sampler (1) and traverse all possible continuous paths from the output of sampler (1) to the input of all samplers. Where these exist, draw in a branch on the signal flow graph from node S_1^* to the respective node S_k^* with the direction and the appropriate sampled transfer function indicated. Then start at sampler (2) and repeat procedure until all samplers are covered.

(6) Now add two other nodes, one labeled with a 1 (the input), and one labeled with a C (the output). From the input node draw a branch to each of the nodes which has a corresponding continuous path in the block diagram, and label with the appropriate gain (RG_{i0} to the output, and the $\overline{RG}_{in}^*$ to the other nodes). Then draw a branch from each of the nodes S_k^* which has a corresponding continuous path in the block diagram to the output node C and label appropriately (G_{n0}').

Then to obtain the output of the system.

(7) Apply Mason's or Coates' method to the nodes marked C and 1 of this sampled-signal-flow graph.

If the sampled form of the output is desired, one may use the standard technique of starring the expression for C obtained in (7).

To illustrate its application, this procedure will be applied to some of the examples given in the paper.

Consider successively the systems of Figs. 1, 4, and 8 of the paper, and in addition the system of Fig. 14. By applying rules 1–6 to these systems, the corresponding sampled signal flow graphs can be drawn directly as shown in Figs. 15, 16, 17, and 18, respectively. By applying Mason's theorem to these flow graphs, one may write by inspection the expressions for the output:

Fig. 15:

$$C = \frac{RG}{1+G} + \frac{\left(\frac{R}{1+G}\right)^* D^*\left(\frac{HG}{1+G}\right)}{1+D^*\left(\frac{HG}{1+G}\right)^*} \quad (43)$$

---

[5] Department of Electrical Engineering, University of California, Berkeley, Calif.

[6] This work was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF 18(600)-1521. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[7] G. G. Lendaris and E. I. Jury, "Input-Output Relationships for Multi-Sampled-Loop Systems," *AIEE Applications and Industry*, January, 1960.

[8] Mason's method referred to in the paper will serve this purpose. It should be pointed out that more recent work has been done in this area by C. L. Coates, "Flow Graph Solutions of Linear Algebraic Equations," *IRE PGCT*, June, 1959.[9] The method he presents is guaranteed to give a minimal term expression for the gain, and for this reason is in many cases superior to Mason's method. However, for the types of flow graphs the control engineer is likely to encounter, the advantage is very little.

[9] For the person who does not wish to pore over the very elegant topological exposition given by Contes, C. A. Desoer has written a very simplified version of the same work with the same results: "The Optimum Formula for the Gain of a Flow-Graph, or, a Simple Derivation of Coates' Formula," to appear in *IRE Proceedings*, March, 1960.

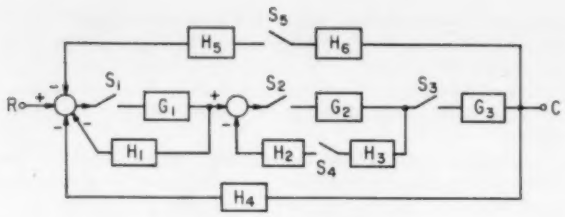


Fig. 14 An $n = 5$ sampled-data system

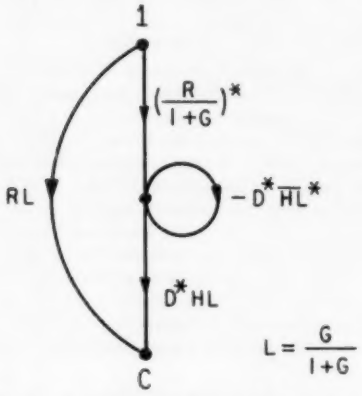


Fig. 15 Sampled-signal-flow graph of Fig. 1

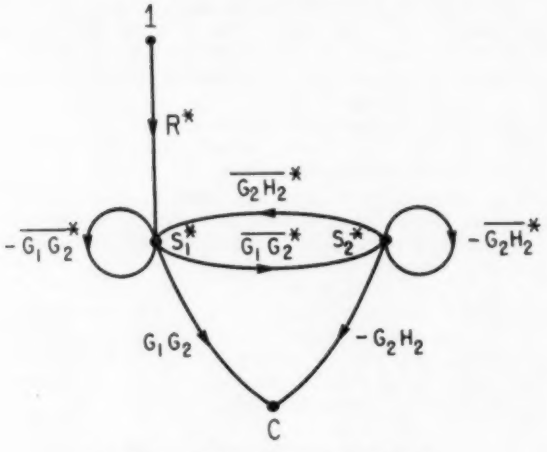


Fig. 16 Sampled-signal-flow graph of Fig. 4

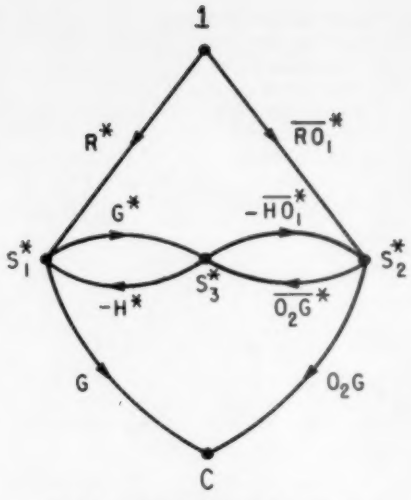


Fig. 17 Sampled-signal-flow graph of Fig. 8

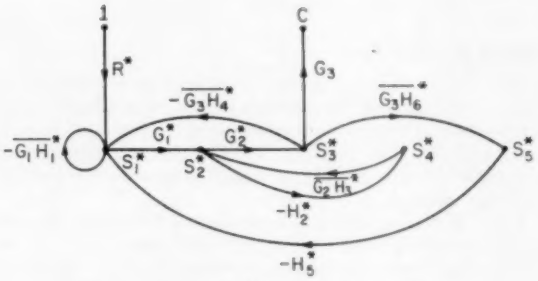


Fig. 18 Sampled-signal-flow graph of Fig. 14

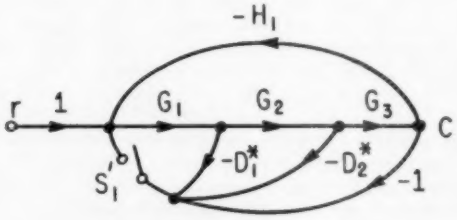


Fig. 19 Reduced form of Fig. 11

Fig. 16:

$$C = \frac{R^*[G_1G_2(1 + \overline{G_2H_2}^*) - \overline{G_1G_2}^*G_2H_2]}{1 + \overline{G_1G_2}^* + \overline{G_2H_2}^*} \tag{44}$$

Fig. 17:

$$C = \frac{R^*G(1 + \overline{HO_1}^*\overline{O_2G}^*) - R^*G^*\overline{HO_1}^*O_2G + \overline{RO_1}^*O_2G(1 + G^*H^*) - \overline{RO_1}^*\overline{O_2G}^*H^*G}{1 + G^*H^* + \overline{HO_1}^*\overline{O_2G}^*} \tag{45}$$

Fig. 18:

$$C = \frac{R^*G_1^*G_2^*G_3}{1 + \overline{G_1H_1}^* + H_2^*\overline{G_2H_3}^* + G_1^*G_2^*\overline{G_2H_4}^*H_5^* + G_1^*G_2^*\overline{G_3H_4}^* + \overline{G_1H_1}^*H_2^*\overline{G_2H_3}^*} \tag{46}$$

The sampled form of Equations (43), (44), and (45) reduce, respectively, to Equations (7), (17), and (22).

Now consider the system of Fig. 11 in the paper. From the sampled signal flow graph of this system, it is noted that this is somewhat of a special system in that none of its determinant terms is zero. Therefore this method has no particular advantage over

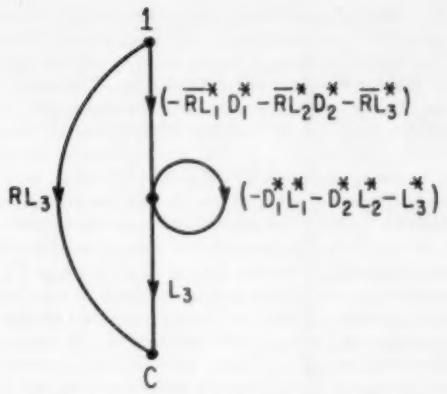


Fig. 20 Sampled signal flow graph of Fig. 19

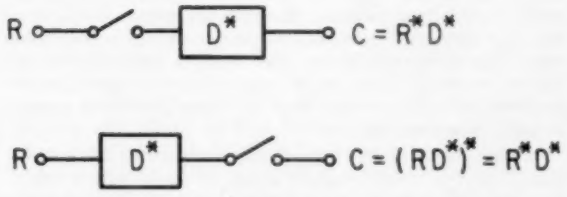


Fig. 21 Effect of (analytically) moving a sampler through a D^* box

that of using the authors' Equation (32), and the continuous version of Equations (36). However, if the analytical trick[10] of moving samplers S_2 and S_3 through their respective D^* boxes is used, there results the one sampler system shown in Fig. 19. Its sampled signal flow graph is given in Fig. 20.

The expression for the output can now be written directly as

$$C = RL_3 - \frac{+(\overline{RL_1}^* D_1^* + \overline{RL_2}^* D_2^* + \overline{RL_3}^*) L_3}{1 + D_1^* L_1^* + D_2^* L_2^* + L_3^*} \tag{47}$$

The sampled form of Equation (47) reduces to Equation (37).

This last example points out that for better efficiency, Rule 1 should be applied to its fullest extent.

Julius T. Tou[11]

The authors have written a very interesting paper on the determination of the output transforms of multiloop sampled systems by use of the signal flow graph techniques. The methods described in this paper are certainly more general and appear to be more useful than the other methods in the literature [2, 3, 5]. The general gain formula, Equation (4), derived in this paper appears quite complicated, and thus is not too convenient to apply. In this respect, the sampled-graph technique seems to be a better method. However, it should be pointed out that to derive the sampled-graph (or pulsed-signal flow graph as called by this discusser) the use of matrix manipulation does not lead to simplification, but it may even complicate the problem which may be readily solved otherwise.

The rules of reduction for conventional signal flow graphs cannot be applied directly to sampled-data systems, because of the presence of samplers in the flow graph. One way of making the conventional rules directly applicable is to reduce the conventional flow graph to one for pulsed signals only. Thus, this method of determination of the output transforms of multiloop sampled-data systems generally involves two steps:

1 Derivation of a pulsed-signal flow graph from the corresponding conventional signal flow graph.

2 Determination of the over-all pulse-transfer function from the pulsed-signal flow graph by use of the rules of flow graph reduction.

The second step is quite straightforward. Clearly, the first step can be easily accomplished when the set of equations

$$x_k = \sum_j \alpha_j x_j^* + \sum_n \beta_n R_n \tag{48}$$

is obtained. In the Equation (48), x_j^* is the pulsed signal from the sampler following node j which will flow to node k; α_j is the transmission from the sampler following node j to node k; R_n is the signal applied at node n which will flow to node k; β_n is the transmission from node n to node k; and x_k is the resulting signal at node k. By applying the principle of superposition, Equation (48) follows immediately by inspection of the conventional signal flow graph describing the sampled-data system. Evidently, no matrix is needed, although one can use it if he likes to. In setting up Equation (48) the use of flow graph reduction rules is sometimes of great help, as demonstrated in an example which follows. Taking the starred transforms (or Z transforms) of both sides of Equation (48) yields

$$x_k^* = \sum_j \alpha_j^* x_j^* + \sum_n \overline{\beta_n R_n}^* \tag{49}$$

The derivation of the pulsed-signal flow graph from Equation (49) is a simple matter. This is the method which is currently used in the digital control course offered at School of Electrical Engineering, Purdue University.

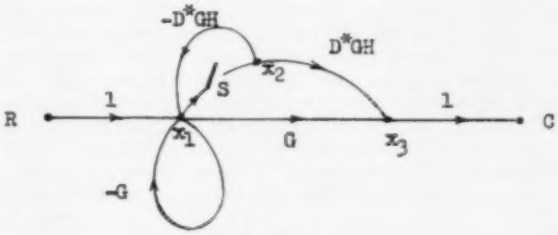


Fig. 22

As an illustration, the sampled-data system of Figs. 1 and 2 of this paper is considered. To facilitate the writing of the equation for x_1, the signal flow graph of Fig. 2 is reduced to the form of Fig. 22 by well-known rules. Inspection of Fig. 22 yields

$$x_1 = -\frac{D^*GH}{1+G} x_1^* + \frac{1}{1+G} R \tag{50}$$

In like manner, the signal flow graph of Fig. 2 is reduced to the form of Fig. 23 in order to facilitate the derivation of the equation for x_3. Thus,

$$x_3 = \frac{D^*GH}{1+G} x_1^* + \frac{G}{1+G} R \tag{51}$$

The equation for x_2 is obvious.

$$x_2 = x_1^* \tag{52}$$

[10] One might argue that physically this step has no meaning, since by its usual definition D^* must have its input in sampled form. However, one may note that analytically the input-output relationships of Fig. 21(*a* and *b*) are both the same:

$$C = R^* D^*$$

[11] Associate Professor, Purdue University, Lafayette, Ind

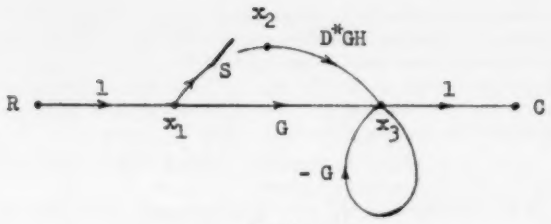


Fig. 23

Taking the starred transforms, these equations are reduced to Equation (6) of the paper from which the pulsed-signal flow graph can readily be drawn. Clearly, following this discussion, a control engineer can derive the pulsed-signal flow graph and then the output transform of a multiloop sampled-data system with ease, even though he is not armed with a sound knowledge of matrix manipulation and topology.

Authors' Closure

The authors wish to thank the discussers for their very interesting and pertinent comments. As Dr. Kuo and Mr. Lendaris point out, the unsampled output can indeed be found without the necessity of matrix multiplication. The application of the "composite" signal flow graph is completely systematic and leads efficiently to a solution. Dr. Kuo's procedure is essentially equivalent to that given by Mr. Lendaris, as may be seen by comparing Figs. 13 and 16. In Fig. 13 we may delete nodes x_2^*, x_3^*, x_4^*, and x_5^*, and also all branches incident on these nodes. Since the nodes in question do not correspond either to the system input, the system output, or to the input of any sampler, their removal has no effect on the computation. The original flow graph may be replaced by the single node labeled c, with branches connecting x_1^* and x_6^*, the weights of the branches being G_1G_2 and $-G_2H_3$, respectively. If this modification is carried out, Fig. 13 becomes identical with Fig. 16. However, it may be desirable to retain the extra nodes in the flow graph; for example, if we wish to consider some other points in the system as the output, we are spared the necessity of drawing another graph.

Professor Tuo suggests that the use of matrix algebra in the derivation of the sampled flow graph may result in complications. The authors emphasize that they have used matrix methods only for the purpose of justifying the final results. In practice, the elements of the sampled flow graph are found not by inversion of matrices, but by applying Mason's formula to compute various transfer functions of the original system, with all samplers open-circuited. It is clear that if the equations describing the system can be manipulated into the form (48), the sampled flow graph may be drawn by inspection. However, the authors do not favor the use of flow graph reduction rules to accomplish the manipulation. The application of the rules requires judgment and intuition; in the analysis of complicated systems it is quite possible to go astray. The authors believe that Mason's formula is more straightforward and efficient.

In response to Professor Tuo's closing remarks, the authors suggest that a control engineer *should* be armed with a sound knowledge of matrix theory and elementary network topology.

R. E. KALMAN
Research Institute for Advanced Study,[2] Baltimore, Md.

J. E. BERTRAM
IBM Research Center, Yorktown Heights, N. Y.

Control System Analysis and Design Via the "Second Method" of Lyapunov[1]

I Continuous-Time Systems

The "second method" of Lyapunov is the most general approach currently in the theory of stability of dynamic systems. After a rigorous exposition of the fundamental concepts of this theory, applications are made to (a) stability of linear stationary, linear nonstationary, and nonlinear systems; (b) estimation of transient behavior; (c) control-system optimization; (d) design of relay servos. The discussion is essentially self-contained, with emphasis on the thorough development of the principal ideas and mathematical tools. Only systems governed by differential equations are treated here. Systems governed by difference equations are the subject of a companion paper.

1—Introduction

One of the most important events in the theory of stability of dynamic systems was the publication in 1892 of Lyapunov's famous memoire in a Russian journal. Translated into French in 1907 and reprinted in America in 1947 [1],[3] it is still relatively unknown and unappreciated in the West. By contrast, the so-called "second method" of Lyapunov has achieved virtual pre-eminence in the Soviet Union as the principal mathematical tool in tackling linear and nonlinear stability problems of the most varied type, particularly in the theory of control systems.

The purpose of this paper is to give a self-contained exposition of the second method of Lyapunov, with particular attention to applications in the theory of control. We shall treat here only continuous-time dynamic systems; i.e., those governed by ordinary differential equations. In a companion paper, an analogous treatment is given of discrete-time dynamic systems.

The objective of the so-called "second method" of Lyapunov is this: *To answer questions of stability of differential equations, utilizing the given form of the equations but without explicit knowledge of the solutions.* The name "second method" is an unfortunate but well-entrenched historical misnomer[4]; actually the "second method" is more accurately described as a point of view, a philosophy of approach, rather than a systematic method. At present, much depends on the ingenuity of the user. In the future, we can hope that systematic procedures will be made possible by machine computation. Yet the concept of a Lyapunov function (see below) appears to be very basic; the whole theory of control systems can be studied in a unified way using Lyapunov functions, as we shall show in what follows. Being a relatively new point of view, the "second method" has much promise for the further development of control theory, particularly as regards nonlinear systems.

The principal idea of the second method is contained in the following physical reasoning: *If the rate of change $dE(\mathbf{x})/dt$ of the energy $E(\mathbf{x})$ of an isolated physical system is negative for every possible state $\mathbf{x}$, except for a single equilibrium state $\mathbf{x}_e$, then the energy will continually decrease until it finally assumes its minimum value $E(\mathbf{x}_e)$.* In other words, a dissipative system perturbed from its equilibrium state will always return to it; this is the intuitive concept of stability.

In order to develop this idea into a precise mathematical tool, the foregoing reasoning must be made independent of the physical concept of energy. As a rule there is no natural way of defining energy when the equations of motion are given in a purely mathematical form. The mathematical counterpart of the foregoing reasoning is the following conjectured theorem: *A dynamic system is stable (in the sense that it returns to equilibrium after any perturbation) if and only if there exists a "Lyapunov function," i.e., some scalar function $V(\mathbf{x})$ of the state with the properties: (a) $V(\mathbf{x}) > 0$, $\dot{V}(\mathbf{x}) < 0$ when $\mathbf{x} \neq \mathbf{x}_e$ and (b) $V(\mathbf{x}) = \dot{V}(\mathbf{x}) = 0$ when $\mathbf{x} = \mathbf{x}_e$.* As will be seen (Theorem 1 and 2), this conjecture is correct under slight additional mathematical restrictions. Occasionally, but not necessarily always, it will suffice to take $V = E$.

One of the attractions of the "second method" is its appeal to geometric intuition. This is illustrated by the following two examples.

Example 1. Harmonic Oscillator. Normalizing the angular frequency to 1, a harmonic oscillator is described by the linear differential equation

$$\ddot{x}_1 + x_1 = 0 \qquad (\dot{} = d/dt)$$

Letting $\dot{x}_1 = x_2$, this equation can be written in the equivalent form

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 \end{aligned}\right\} \tag{1}$$

The state of the system is represented by the vector $\mathbf{x} = (x_1, x_2)$. The solutions (trajectories) of (1) in the state-plane (phase-plane) are concentric circles about the origin, Fig. 1. The energy of the system is given by

$$E(\mathbf{x}) = E(x_1, x_2) = x_1^2 + x_2^2 = V(\mathbf{x}) \tag{2}$$

The derivative $\dot{V}$ of V along any solution of (1), i.e., along any trajectory in Fig. 1, is given by

---

[1] This research was supported in part by the U. S. Air Force Office of Scientific Research under Contract AF 49(638)-382.

[2] 7212 Bellona Ave.

[3] Numbers in brackets designate References at end of paper.

[4] The so-called "first method" of Lyapunov deals with stability questions via an explicit representation of the solutions of a differential equation.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE-IRE-ISA National Automatic Control Conference, Dallas, Tex., November 4–5, 1959.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 10, 1959. Paper No. 59—NAC-2.

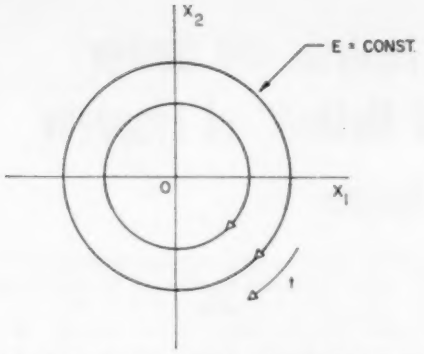


Fig. 1 Phase-plane trajectories of harmonic oscillator

$$\dot{V}(\mathbf{x}) = \frac{dV(\mathbf{x}(t))}{dt} = \dot{\mathbf{x}}'[\text{grad } V(\mathbf{x})]$$

$$= \dot{x}_1 \frac{\partial V}{\partial x_1} + \dot{x}_2 \frac{\partial V}{\partial x_2}$$

$$= 2\dot{x}_1 x_1 + 2\dot{x}_2 x_2$$

which is identically zero by (1). Hence $\dot{V}(\mathbf{x}) = 0$; the energy remains constant—(1) is a conservative system.

Example 2. Asymptotically Stable Nonlinear System. Consider now a slight modification of system (1):

$$\left.\begin{aligned} \dot{x}_1 &= x_2 - ax_1(x_1^2 + x_2^2) \\ \dot{x}_2 &= -x_1 - ax_2(x_1^2 + x_2^2) \end{aligned}\right\} \tag{3}$$

where a is a positive constant.

We again define V by (2), but now of course this quantity is not necessarily the energy of system (3). Calculating as before, the derivative $\dot{V}$ of V along a solution of (3) is

$$\dot{V}(\mathbf{x}) = \frac{dV(\mathbf{x}(t))}{dt} = -2a(x_1^2 + x_2^2)^2 \tag{4}$$

$$= -2aV^2(\mathbf{x})$$

which is negative unless $x_1 = x_2 = 0$. This shows that V is constantly decreasing along any solution of (3), and hence V is a Lyapunov function.

With reference to Fig. 2, it follows also that the trajectories of (3) cross the boundary of every region $V(\mathbf{x}) \leqq$ const from the outside toward the inside. From this we are led to the following geometric interpretation of a Lyapunov function:

Let $V(\mathbf{x})$ be a measure of the "distance" of the state $\mathbf{x}$ from the origin in the state space, i.e., $V(\mathbf{x}) > 0$ when $\mathbf{x} \neq \mathbf{0}$ and $V(\mathbf{0}) = 0$. Suppose the distance between the origin and the instantaneous state $\mathbf{x}(t)$ is continually decreasing as $t \to \infty$, i.e., $\dot{V}(\mathbf{x}(t)) < 0$. Then $\mathbf{x}(t) \to \mathbf{0}$.

In Examples 1–2, $V(\mathbf{x})$ is actually the square of the Euclidean distance. In general $V(\mathbf{x}) =$ const turns out to be (because of $\dot{V}(\mathbf{x}) < 0$) a family of concentric closed surfaces surrounding the origin such that the surface $V(\mathbf{x}) = c_1$ lies inside the surface $V(\mathbf{x}) = c_2$ whenever $c_1 < c_2$. Thus we can say that $\mathbf{y}$ is "farther" away from the origin than $\mathbf{x}$ if $V(\mathbf{y}) > V(\mathbf{x})$.

Unlike the energy of a system, the Lyapunov function $V(\mathbf{x})$ is not unique; this is precisely the reason why the "second method" of Lyapunov is a more powerful tool than conventional energy considerations. A system whose energy E decreases *on the average*, but not necessarily at each instant, is stable but E is not a Lyapunov function.

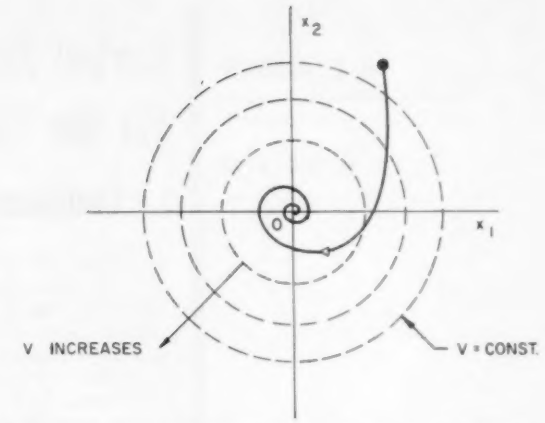


Fig. 2 Phase-plane trajectories of harmonic oscillator with nonlinear damping

2—Outline of Contents

Sections 4–6 together with the Appendix provide the necessary mathematical machinery. Sections 7–10 contain the applications.

Description of Dynamic Systems (Section 4). To apply the "second method," it is necessary to describe dynamic systems from the "state" point of view. Since this may be unfamiliar to many control engineers, an outline of the most important concepts and definitions is provided.

Definitions of Stability (Section 5). In treating general types of systems (nonlinear, nonstationary) it is necessary to make careful distinctions between various intuitive aspects of stability. The necessary definitions are stated precisely and in accordance with current mathematical usage.

The Main Theorems (Section 6). Theorem 1 gives the conditions on a Lyapunov function whose existence guarantees the strongest type of stability. In the corollaries of Theorem 1, these conditions are stated in various weakened forms suitable for certain applications. Theorem 3 summarizes the current state of knowledge concerning the stability of linear, nonstationary systems; Corollaries 3.1–2 provide a purely algebraic procedure for solving many problems in linear stationary systems, which is simpler than the customary methods based on the Laplace transform and is well suited for machine computation. All proofs of this section may be omitted at first reading without loss of continuity.

Applications to Stability Theory (Section 7). An easy consequence of Corollary 3.1 in the linear, stationary case is a rigorous elementary derivation of the Routh-Hurwitz inequalities (Example 6). The paper and pencil derivation is very messy, but the method lends itself well to machine computation. Example 7 gives the construction of a Lyapunov function for the canonic form of feedback systems. Example 8 is an application of the "second method" to an electric circuit consisting of a time-varying resistor, inductor, and capacitor. Example 9 gives a rigorous justification of the usual linearization procedure. Example 10 is a proof of the stability of passive nonlinear networks. Examples 11 and 13 give conditions, similar to the Routh-Hurwitz inequalities, for the stability of some nonlinear systems; Theorem 4 is probably the best available result along these lines which does not depend on the order of the system.

Estimation of Transient Behavior (Section 8). Since a Lyapunov function can be regarded as defining distance from the origin in the state space, its derivative $\dot{V}$ can be used to give a quantitative estimate of the speed with which the origin is approached. This

leads to an equivalent time constant even for nonlinear systems (see Example 14). Similarly, the Lyapunov function can be used to estimate the effect of random perturbations and sensitivity to parameter variations, as in Example 15.

Relations With System Optimization (Section 9). If a performance index is defined as the error criterion integrated with respect to time, then the performance index is actually a Lyapunov function—provided the error criterion is not identically zero along any trajectory of the system. This observation (Theorem 5) removes a serious shortcoming of the conventional Wiener theory of quadratic optimization since the latter does not automatically guarantee that the quadratic optimal system is stable. Example 16 shows how a conventional quadratic optimization problem can be solved by means of Corollary 3.1; Example 17 is a rigorous treatment of the same problem using elementary concepts from game theory.

Design of a Relay Servo (Section 10). While exact optimization is usually difficult or impractical, the "second method" can supply some new ideas for approximate optimization. In Example 18 we present a very simple design algorithm of this type for a stable relay servo (again based on Corollary 3.1) which requires only a linear compensating element. The analytical findings are supported by an analog-computer study.

3—Guide to Western Literature

There are very few expositions of the subject in Western languages; most of those available are written from the abstract mathematical point of view. The following comments are intended to facilitate study of the subject from the engineering point of view.

The original work of Lyapunov [1] (in French) is difficult to read because of obsolescent mathematical terminology. An authoritative survey of mathematical results until 1956 is given by Massera [2]. The recent monograph of Hahn [3] (in German) covers existing theoretical results and includes a few applications. There is now an English translation [4] of the well-known book of Malkin. This book, allegedly addressed to engineers, is primarily a detailed and rigorous mathematical treatment of classical problems. There is a section (Chapter VI, par. 3) on the basic theorems of Lyapunov in the well-known book of Lefschetz [5]; see also his paper [6] on recent generalizations. The problem of control-system design has been treated at length by Lur'e [7] (German and English translation) and Letov [8].

Hahn [3] and the survey article of Antosiewicz [9] give many references to Soviet literature. Over the past 10 years, a great deal has been published about the "second method" in the journals *Avtomatika i Telemekhanika* and *Prikladnaya Matematika i Mekhanika* (now both available in English translation).

4—Description of Continuous-Time Dynamic Systems

From now on extensive use will be made of vector notation. Small boldface Roman or Greek letters will denote vectors. Capital boldface Roman or Greek letters will denote matrices. The unit matrix is $\mathbf{I}$, the zero vector or matrix $\mathbf{0}$. The transpose of a vector or matrix is denoted by the prime; the scalar product of $\mathbf{x}$, $\mathbf{y}$ is $\mathbf{x}'\mathbf{y}$, and the quadratic form associated with a square matrix $\mathbf{A}$ is $\mathbf{x}'\mathbf{A}\mathbf{x}$. The norm is denoted by $\|\mathbf{x}\|$; in specific calculations this is to be taken as the Euclidean norm $(\mathbf{x}'\mathbf{x})^{1/2}$. The eigenvalues of a matrix $\mathbf{F}$ are $\lambda_i(\mathbf{F})$.

In this paper we shall study systems governed by the *vector differential equation:*

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, \mathbf{u}(t), t), \qquad -\infty < t < +\infty \tag{5}$$

This is equivalent to the set of n scalar differential equations

$$dx_i/dt = f_i(x_1, \ldots, x_n, u_1(t), \ldots, u_m(t), t) \qquad (5)_i$$
$$(i = 1, \ldots, n)$$

The vector $\mathbf{x}$ is the *state* of the system (5), its components x_i are the *state variables*. The vector $\mathbf{u}(t)$ is the *control* function (or *forcing function* or *input*) of the system (5), its components $u_i(t)$ are the *control variables*. The system is specified by the vector-valued function $\mathbf{f}$. The integer n is the *order* of the system.

If $\mathbf{u}(t) \equiv \mathbf{0}$ for all t, we say that (5) is *free* (*unforced*):

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, t) \tag{5-F}$$

It is always assumed that the function $\mathbf{f}$ in (5) is sufficiently smooth so that the equation has a unique solution starting at any initial state $\mathbf{x}_0$ at any time t_0.

To make this idea precise, we first fix $\mathbf{u}(t)$ in (5), so that we might as well consider (5-F). We assume that *there exists a unique vector function* $\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)$, *differentiable in* t, *such that, for any fixed* $\mathbf{x}_0$, t_0

(i) $\boldsymbol{\phi}(t_0;\ \mathbf{x}_0, t_0) = \mathbf{x}_0$
(ii) $d\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)/dt = \mathbf{f}(\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0), t)$

in some interval $|t - t_0| \leq a(t_0)$. The function $\boldsymbol{\phi}$ is called a *solution* of the system (5-F) or (5).

In older texts on differential equations, particularly in the engineering literature, it is customary to write $\mathbf{x}(t)$ for the solution. This "informal" usage will be employed here only rarely; reasons of clarity demand in most cases (as in Massera [2] and Hahn [3]) the use of the explicit notation just introduced. Thus $\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)$ means: a solution of (5), with fixed $\mathbf{u}(t)$, going through state $\mathbf{x}_0$ at time t_0, observed at time t. We also call $\boldsymbol{\phi}$ the *transition function* since it shows how $\mathbf{x}(t_0)$ is transformed into $\mathbf{x}(t)$. With this interpretation, $\mathbf{f}$ in (5) may be called the *infinitesimal transition function* since the differential equation specifies the infinitesimal state transition $\mathbf{x} \to \mathbf{x} + d\mathbf{x}$ corresponding to the infinitesimal change in time $t \to t + dt$.

Now if the differential equation (5) is to represent a physical phenomenon, one would always assume also that $\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)$ *is continuous in all of its arguments* (of course, continuity in t is already demanded by (ii) above!). A sufficient condition for the validity of the foregoing assumptions is provided by the following classical theorem (Ref. [11], Chap. 1, Theorems 2.3 and 7.1):

EXISTENCE, UNIQUENESS, AND CONTINUITY THEOREM. *Let* $\mathbf{f}(\mathbf{x}, t)$ *be continuous in* $\mathbf{x}$, t, *and satisfy a Lipschitz condition in some region about any* $\mathbf{x}_0$, t_0:

$$R(\mathbf{x}_0, t_0) = \begin{cases} \|\mathbf{x} - \mathbf{x}_0\| \leq b(\mathbf{x}_0) \\ |t - t_0| \leq c(t_0) \end{cases} \qquad (b, c > 0)$$

i.e., the following condition is satisfied for $(\mathbf{x}, t)$, $(\mathbf{y}, t)$ *in* $R(\mathbf{x}_0, t_0)$

$$\|\mathbf{f}(\mathbf{x}, t) - \mathbf{f}(\mathbf{y}, t)\| \leq k\|\mathbf{x} - \mathbf{y}\|$$

where k *is a positive constant which depends only on* b, c. THEN

1 *There exists a unique solution* $\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)$ *of* (5-F) *starting at* $\mathbf{x}_0$, t_0, for all $|t - t_0| \leq a(t_0)$,

$$a(t_0) \geq \text{Min}\{c(t_0), b(\mathbf{x}_0)/M(\mathbf{x}_0, t_0)\}$$

where $M(\mathbf{x}_0, t_0)$ *is the maximum assumed by the continuous function* $\|\mathbf{f}(\mathbf{x}, t)\|$ *in the closed, bounded set* $R(\mathbf{x}_0, t_0)$.

2 *In some small neighborhood of* $\mathbf{x}_0$, t_0 *the solution is a continuous function of its arguments.*

Note that the Lipschitz condition implies continuity of $\mathbf{f}$ in $\mathbf{x}$, but not necessarily in t; it is implied by bounded partial derivatives in $\mathbf{x}$. It should be borne in mind that the *local* Lipschitz condition required by the theorem only implies the desired properties of a solution *near* $\mathbf{x}_0$, t_0. From it we cannot conclude the

existence of solutions for arbitrarily large values of t as the following example shows:

Example 3. Finite Escape Time. Consider the first-order nonlinear differential equation

$$\dot{x}_1 = x_1^2$$

The requirements of the existence and uniqueness theorem are obviously satisfied.

By separation of variables, the solution is seen to be

$$-\frac{1}{\phi_1(t;\ x_{01},\ t_0)} + \frac{1}{x_{01}} = t - t_0$$

Now if $x_{01} > 0$, the solution will increase with t; therefore $\phi_1(t;\ x_{01},\ t_0)$ tends to infinity as t tends to $t_1 = t_0 + 1/x_{01}$. As t increases beyond t_1, the original solution which started at t_0, $x_{01} > 0$ cannot be continued, except possibly by letting the state $x_1(t_1)$ "jump" from $+\infty$ to $-\infty$, which is clearly not the way physical systems behave.

Following Markus [17], one says in this case that the solution of the differential equation escapes to infinity in finite time or has *finite escape time*. (This can be always avoided by replacing t by $t' = t/(t_\infty - t)$.) In order that a differential equation represent a physical system, this *possibility must be ruled out by explicit assumption to the contrary.*

If the Lipschitz condition holds for $\mathbf{f}$ *everywhere*, there can be no finite escape time. To see this (it does not follow directly from the theorem!), let

$$\|\mathbf{f}(\mathbf{x},\ t) - \mathbf{f}(\mathbf{y},\ t)\| \leq k\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}$, $\mathbf{y}$, t.

Integrating both sides of (5-F) and using this, and (8-A)

$$\|\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0)\| \leq \|\mathbf{x}_0\| + \left\|\int_{t_0}^{t} \mathbf{f}(\boldsymbol{\phi}(\tau;\ \mathbf{x}_0,\ t_0),\ \tau)d\tau\right\|$$

$$\leq \|\mathbf{x}_0\| + k\int_{t_0}^{t} \|\boldsymbol{\phi}(\tau;\ \mathbf{x}_0,\ t_0)\|d\tau$$

This is easily seen to imply, by the well-known Gronwall-Bellman lemma (Reference [49], p. 35), that

$$\|\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0)\| \leq [\exp k(t - t_0)]\|\mathbf{x}_0\| \tag{6}$$

which is less than infinity for any finite $(t - t_0)$.

However, for some purposes the requirement of a *global* Lipschitz condition is too restrictive; for instance, in case of servos with ramp inputs.

If the foregoing hypotheses concerning $\boldsymbol{\phi}$ are satisfied, the differential equation (5) is said to represent a *continuous-time dynamic system*. We then call $\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0)$ a *motion* (or *trajectory* or *transition function*) of the dynamic system. Frequently, it is convenient to define a dynamic system axiomatically by means of its motions satisfying the foregoing hypotheses [10]. These axioms may be stated as follows:

(A_1) (i) $\boldsymbol{\phi}(t_0;\ \mathbf{x}_0,\ t_0) = \mathbf{x}_0$ for all $\mathbf{x}_0,\ t_0$ (7)

(ii) $\boldsymbol{\phi}(t_2;\ \boldsymbol{\phi}(t_1;\ \mathbf{x}_0,\ t_0),\ t_1)$
$= \boldsymbol{\phi}(t_2;\ \mathbf{x}_0,\ t_0)$ for all $\mathbf{x}_0,\ t_0,\ t_1,\ t_2$ (8)

(*existence* and *uniqueness;* see Fig. 3.)

(A_2) $\boldsymbol{\phi}$ *is continuous with respect to all arguments.*
(A_3) $\boldsymbol{\phi}$ *is defined for all* $\mathbf{x}_0,\ t_0,\ t$ (no finite escape time).

A state $\mathbf{x}_e$ of a free dynamic system (5-F) is an *equilibrium state* if

$$\mathbf{f}(\mathbf{x}_e;\ t) = \mathbf{0} \quad \text{for all} \quad t$$

or, equivalently,

$$\boldsymbol{\phi}(t;\ \mathbf{x}_e,\ 0) = \mathbf{x}_e \quad \text{for all} \quad t$$

In other words, a motion passing through an equilibrium state at any time is actually at the same state at all times.

A dynamic system (5) is *stationary* if

$$\mathbf{f}(\mathbf{x},\ \mathbf{u}(t),\ t) \equiv \mathbf{f}(\mathbf{x},\ \mathbf{u}(t)) \qquad \text{for all} \qquad \mathbf{x},\ \mathbf{u}(t)$$

i.e., if $\mathbf{f}$ does not depend explicitly on time (but may depend implicitly on time through $\mathbf{u}(t)$). For a free system, this is equivalent to saying that every motion is *invariant under translation in time*, i.e., that

$$\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0) = \boldsymbol{\phi}(t + \tau;\ \mathbf{x}_0,\ t_0 + \tau)$$

for any $\mathbf{x}_0$, t_0, τ. A system which is both free and stationary is said to be *autonomous.*

A dynamic system (5) is *linear* if $\mathbf{f}$ is a linear function of $\mathbf{x}$ and $\mathbf{u}$. Assuming for simplicity that $\mathbf{0}$ is an equilibrium state, i.e., that

$$\mathbf{f}(\mathbf{0},\ \mathbf{0},\ t) = \mathbf{0} \tag{8-A}$$

it follows that a linear system (5) can be written in the form

$$d\mathbf{x}/dt = \mathbf{F}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}(t) \tag{5-L}$$

where

$$\mathbf{F}(t) = [\partial f_i(t)/\partial x_j]$$

$$\mathbf{D}(t) = [\partial f_i(t)/\partial u_j]$$

are $n \times n$ resp. $n \times m$ matrices depending on time. In a stationary linear system $\mathbf{F}$ and $\mathbf{D}$ are constant matrices.

For $t \geq t_0$, the general solution of (5-L) can be written as [11]

$$\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0) = \boldsymbol{\Phi}(t,\ t_0)\mathbf{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t,\tau)\,\mathbf{D}(\tau)\mathbf{u}(\tau)d\tau \tag{9}$$

where $\boldsymbol{\Phi}(t,\ t_0)$ is called the *transition matrix* of the system (5-L). Its elements may be regarded as the impulse responses under appropriate excitations and observations [12, 13]. Thus $\phi_{ij}(t,\ t_0)$ is the response of the ith state variable to an excitation defined by (taking $\mathbf{D}(t) \equiv \mathbf{I}$):

$$u_j(t) = \delta(t - t_0);\ u_k(t) = 0 \quad \text{when} \quad k \neq j$$

with $\mathbf{x}_0 = \mathbf{0}$.

As a consequence of linearity, relations (7) and (8) now imply (set $\mathbf{u}(t) \equiv \mathbf{0}$ in (9)!)

$$\boldsymbol{\Phi}(t_0,\ t_0) = \mathbf{I} \qquad \text{for all } t_0 \tag{7-L}$$

and

$$\boldsymbol{\Phi}(t_2,\ t_0) = \boldsymbol{\Phi}(t_2,\ t_1)\boldsymbol{\Phi}(t_1,\ t_0) \quad \text{for all} \quad t_0,\ t_1,\ t_2 \tag{8-L}$$

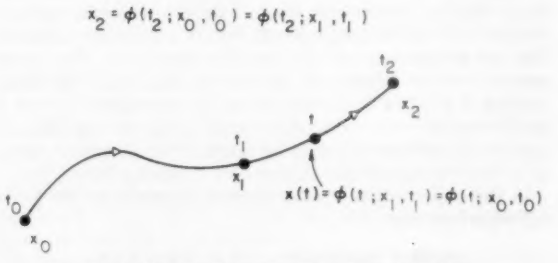


Fig. 3 Notation for motions of a dynamic system

From this, it follows at once also that

$$\Phi^{-1}(t, t_0) = \Phi(t_0, t) \qquad \text{for all } t, t_0 \tag{9}$$

Another useful property of the transition matrix, an immediate consequence of its definition by (9), is that it satisfies its own differential equation:

$$d\Phi(t, t_0)/dt = \mathbf{F}(t)\Phi(t, t_0) \quad \text{for all } t, t_0 \tag{10}$$[5]

In a stationary linear system $\Phi(t, t_0)$ depends only on the difference $t - t_0$. In fact, it is easy to show [5, 11] that

$$\Phi(t, t_0) = \sum_{k=0}^{\infty} [(t - t_0)\mathbf{F}]^k/k! = \exp(t - t_0)\mathbf{F} \tag{11}$$

As usually stated, the "second method" of Lyapunov applies only to free dynamic systems, with equilibrium state at $\mathbf{0}$; that is to say, to deviations about some fixed motion. To see this, let $\mathbf{y}(t)$ be some fixed motion of (5), corresponding to a fixed choice of $\mathbf{u}(t)$. Let

$$\mathbf{x} = \mathbf{y}(t) + \mathbf{z}$$

Then

$$\begin{aligned} d\mathbf{x}/dt &= d\mathbf{y}(t)/dt + d\mathbf{z}/dt \\ &= \mathbf{f}(\mathbf{y}(t) + \mathbf{z}, \mathbf{u}(t), t) \\ &= \mathbf{f}(\mathbf{y}(t), \mathbf{u}(t), t) + \mathbf{g}(\mathbf{z}, t) \end{aligned}$$

Since $\mathbf{y}(t)$ is a solution of (5), it follows that

$$\begin{aligned} &d\mathbf{z}/dt = \mathbf{g}(\mathbf{z}, t), \\ &\mathbf{g}(\mathbf{0}, t) = \mathbf{0} \quad \text{for all } t \end{aligned}$$

so that, for fixed $\mathbf{u}(t)$, deviations from the motion $\mathbf{y}(t)$ are a free dynamic system.

The description of dynamic systems via the concept of "state" (instead of more restricted concepts like impulse response, frequency response, transfer function, etc., which are invalid for nonlinear and sometimes even for nonstationary linear systems) is essential for the understanding of the "second method." In fact, this is necessary even for a precise definition of stability. Though not yet in common use, the "state" method has been applied recently to numerous control problems; see references [12–16] which also contain much pedagogical material in regard to setting up of equations, interpretation of solutions in geometric terms, etc. The idea of "state" is well known, of course; it has been used in mathematics more or less explicitly at least since Poincaré.

5—Concepts of Stability

On closer examination, the intuitive idea of stability turns out to be very complex. This is particularly true in nonlinear dynamic systems. As a result, a large variety of precise definitions of stability have been proposed, often differing in subtle ways which are confusing to the uninitiated. Here we shall present the principal definitions used at present and discuss their relationships to one another.

The question of stability concerns deviation about some *fixed* motion. By the remarks at the end of the previous section, it follows that it suffices to consider deviations from an equilibrium state $\mathbf{x}_e$ of a free dynamic system.

Perhaps the oldest concept of stability is this: If the system (5-F) is perturbed slightly from its equilibrium state at the origin, all subsequent motions remain in a correspondingly small neighborhood of the origin. For instance, harmonic oscillators (Example 1) possess this type of stability. The precise definition goes back to Lyapunov:

(S_1) An equilibrium state $\mathbf{x}_e$ of a free dynamic system is *stable* if for every real number $\epsilon > 0$ there exists a real number $\delta(\epsilon, t_0) > 0$ such that $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq \delta$ implies

$$\|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq \epsilon \qquad \text{for all } t \geq t_0$$

Fig. 4 illustrates the definition. Note that stability in the sense of Lyapunov (definition (S_1)) is a *local* concept; it refers to behavior near the equilibrium state since one does not know *a priori* how small δ in the definition may have to be chosen.

Observe further that *if there is stability for some initial time t_0 there is stability for any other initial time t_1, provided all motions are continuous in the initial state* (Axiom (A_3)).

In fact, if $t_1 < t_0$ define (possible, since ϕ is continuous in t')

$$\Omega(\mathbf{x}_1, t_1, t_0) = \operatorname*{Max}_{t_1 \leq t \leq t_0} \|\phi(t;\ \mathbf{x}_1, t_1) - \mathbf{x}_e\| \tag{12}$$

This function is clearly 0 when $\mathbf{x}_1 = \mathbf{x}_e$; elementary but somewhat tedious arguments show that, due to (A_2), Ω is continuous in $\mathbf{x}_1$. Hence there is some $\nu(\delta(\epsilon, t_0), t_1) > 0$ such that $\|\mathbf{x}_1 - \mathbf{x}_e\| \leq \nu$ implies $\|\phi(t;\ \mathbf{x}_1, t_0) - \mathbf{x}_e\| \leq \delta$ in the interval $t_1 \leq t \leq t_0$. By stability at t_0, it follows that $\|\phi(t;\ \mathbf{x}_1, t_1) - \mathbf{x}_e\| \leq \epsilon$ for all $t \geq t_1$. We can take $\delta(\epsilon, t_1) = \nu$. If $t_0 \leq t_1$, merely let $\Omega = \|\phi(t_0;\ \mathbf{x}_1, t_1)\|$.

In most engineering applications, one is interested in a stronger kind of stability; namely, the motion should return to equilibrium after any small perturbation. The classical definition of Lyapunov is the following:

(S_2) An equilibrium state $\mathbf{x}_e$ of a free dynamic system is *asymptotically stable* if

(i) it is stable and

(ii) every motion starting sufficiently near $\mathbf{x}_e$ converges to $\mathbf{x}_e$ as $t \to \infty$. In other words, there is some real constant $r(t_0) > 0$ and to every real number $\mu > 0$ there corresponds a real number $T(\mu, \mathbf{x}_0, t_0)$ such that $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq r(t_0)$ implies

$$\|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq \mu \text{ for all } t \geq t_0 + T$$

See Fig. 5. Asymptotic stability is also a local concept; one does not know *a priori* how small $r(t_0)$ may have to be. If a free *linear* system is asymptotically stable, then $\|\mathbf{x}_1 - \mathbf{x}_e\| < \|\mathbf{x}_0 - \mathbf{x}_e\|$ implies that $T(\mu, \mathbf{x}_1, t_0) < T(\mu, \mathbf{x}_0, t_0)$ for any μ and t_0, since the initial state is factored out in ϕ (see (9)). In other words, among motions starting at the same *distance* from $\mathbf{x}_e$ none will remain at a distance larger than μ from $\mathbf{x}_e$ at arbitrarily large values of time. This is expressed by the following definition due to Massera [18]:

(S_3) An equilibrium state $\mathbf{x}_e$ of a free dynamic system is *equiasymptotically stable* if

(i) it is stable and

(ii) every motion starting sufficiently near $\mathbf{x}_e$ converges to $\mathbf{x}_e$ as $t \to \infty$ uniformly in $\mathbf{x}_0$. In other words, T in definition (S_2-ii) is of the form $T(\mu, r(t_0), t_0)$.

A somewhat puzzling aesthetic deficiency of definitions (S_2 – S_3) is that one has to assume *explicitly* that (S_1) holds. The following example shows that (S_2-ii) does not imply (S_1) or (S_3-ii) when $n > 1$.

Example. Consider the second-order system in polar co-ordinates $\mathbf{x} = (r, \theta)$, $0 \leq r < \infty$, $0 \leq \theta < 2\pi$:

$$\left.\begin{aligned} \dot{r} &= [g(\theta, t)/g(\theta, t)]r \\ \dot{\theta} &= 0 \end{aligned}\right\}$$

[5] In the literature [5, 11], any nonsingular matrix function of time satisfying (10) is called a *fundamental matrix* of (5-L). In our case, the transition matrix is a fundamental matrix made *unique* through the requirement (7-L).

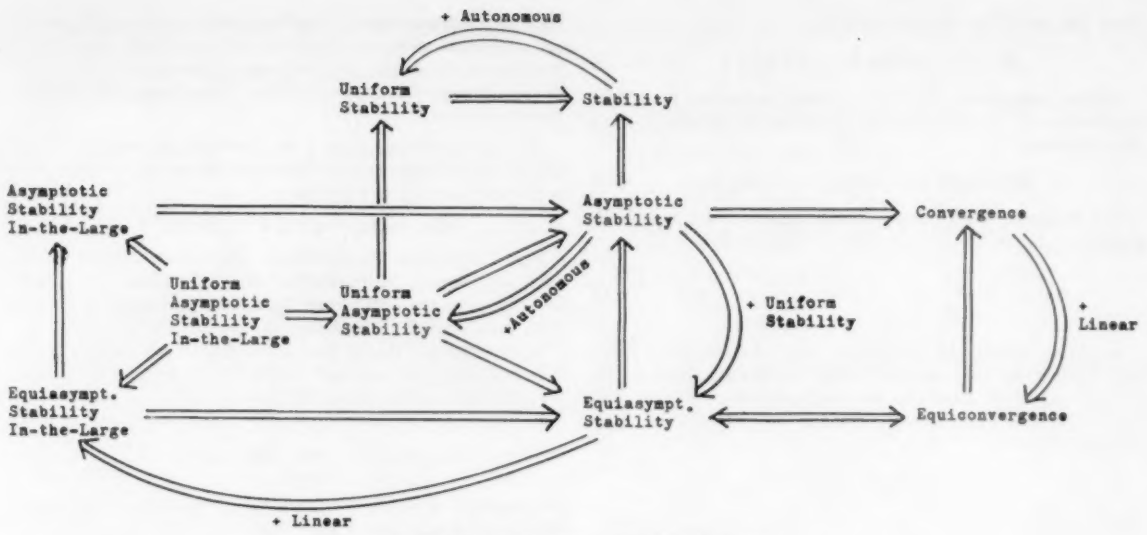


Fig. 6 Interrelations between stability concepts

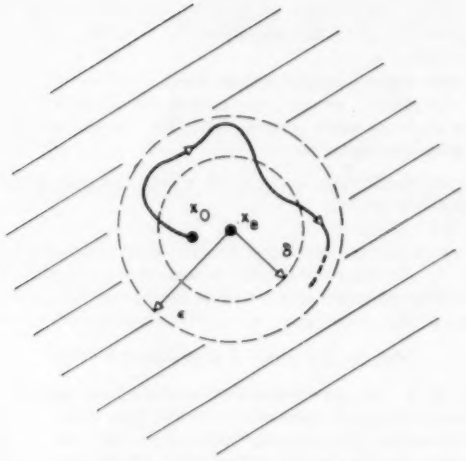


Fig. 4 Definition of stability

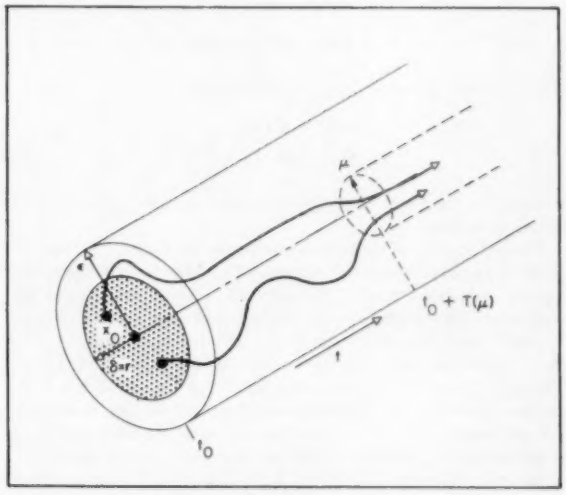


Fig. 5 Definition of asymptotic stability

where

$$g(\theta, t) = \frac{\sin^2 \theta}{\sin^4 \theta + (1 - t \sin^2 \theta)^2} + \frac{1}{1 + t^2}$$

Then

$$r(t;\ r_0, \theta_0, t_0) = [g(\theta_0, t)/g(\theta_0, t_0)]r_0$$

$$\theta(t;\ r_0, \theta_0, t_0) = \theta_0$$

The motion is continuous with respect to r_0, θ_0. (S_T-ii) is satisfied. However, (S_1), (S_2-ii) do not hold because at $t_1 = \sin^{-2}\theta_0$, $g(\theta_0, t_1) > \sin^{-2}\theta_0$, which can be made arbitrarily large as $\theta_0 \to \pm\pi$. Here the function $T(\mu, \mathbf{x}_0, t_0)$ is not continuous in $\mathbf{x}_0$ along the ray $\theta_0 = \pm\pi$.

On the other hand, (S_T-ii) *implies* (S_1) *if all motions are continuous in* $\mathbf{x}_0$.

In fact, let $t_1 = t_0 + T(\epsilon, r(t_0), t_0)$ and define

$$\Lambda(\mathbf{x}_0, t_0, T) = \operatorname*{Max}_{t_0 \le t \le t_1} \|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\|$$

Since T does *not* depend on $\mathbf{x}_0$ by (S_T-ii), it follows by (A_3), as in (12), that Λ is continuous in $\mathbf{x}_0$. Hence there is some $\delta(\epsilon, t_0)$ such that $\|\mathbf{x}_0 - \mathbf{x}_e\| \le \delta$ implies

$$\|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\| \le \epsilon \text{ for all } t_0 \le t \le t_0 + T$$

By (S_T-ii), the inequality then holds for all $t \ge t_0$.

(S_3) obviously implies (S_2). (S_2) *is equivalent to* (S_3) *if the dynamic system is of the first order* [18]. In fact, if $x_e < y < x$ (scalars), then for all t, t_0 $\phi(t;\ y, t_0) < \phi\ (t; x, t_0)$, for otherwise by continuity in t one would have $z = \phi(t_1;\ x, t_0) = \phi(t_1;\ y, t_0)$ which says that the motion starting at x, t_1 may be either at x or y at t_0, contradicting axiom (A_1). Therefore, if

$$|x| \le r(t_0),\ T(\mu, x_0, t_0) \le \text{Max } T(\mu, \pm\, r(t_0), t_0)$$

Another important property of linear systems is that stability is independent of the distance of the initial state from $\mathbf{x}_e$. This is expressed by

(S_4) an equilibrium state $\mathbf{x}_e$ of a free dynamic system is *asymptotically* [(S_4): *equiasymptotically*] *stable in the large* if

(i) it is stable and

(ii) every motion converges to $\mathbf{x}_e$ [converges to $\mathbf{x}_e$ uniformly in $\mathbf{x}_0$ for $\|\mathbf{x}_0\| \leq r$ where r is fixed but arbitrarily large] as $t \to \infty$.

If a system is linear (S_2-S_4) are equivalent.

A motion is said to be *bounded* for every $\mathbf{x}_0$, t_0 if there is some constant $B(\mathbf{x}_0, t_0)$ such that $\|\phi(t;\ \mathbf{x}_0, t_0)\| \leq B$ for all $t \geq t_0$. The motion is *equibounded* if $B(\mathbf{x}_0, t_0) \leq B(r, t_0)$ for all $\|\mathbf{x}_0\| \leq r$. These properties are implied by asymptotic resp. equiasymptotic stability in the large as is seen from the proof that (S_2-ii) implies (S_3).

If a system is autonomous (= free + stationary), then δ, r, T in the preceding definitions do not depend on t_0. It is not hard to show [2] that the same holds also if the function $\mathbf{f}(\mathbf{x}, t)$ is periodic in t. There are systems which have this property without being autonomous or periodic. This motivates the various definitions of *uniform* (in time) stability.

(S_6) *Uniform stability* is stability such that δ does not depend on t_0.

It turns out that *asymptotic and equiasymptotic stability are equivalent if they are uniform in* t_0.

In fact, take any $\mu > 0$. Since the foregoing hypotheses imply (S_6), there is some corresponding $\delta(\mu) > 0$. Consider any $\mathbf{x}$ in the region $\|\mathbf{x} - \mathbf{x}_e\| \leq r$. By uniform (in t_0) convergence, find a $T(\delta/2, \mathbf{x})$ such that

$$\|\phi(t;\ \mathbf{x}, t_0) - \mathbf{x}_e\| \leq \delta/2 \quad \text{for all } t \geq t_0 + T$$

By continuity with respect to the initial state, there is some neighborhood $U(\mathbf{x})$ of $\mathbf{x}$ such that if $\mathbf{y}$ is in $U(\mathbf{x})$ then

$$\|\phi(t_0 + T;\ \mathbf{y}, t_0) - \mathbf{x}_e\| \leq \delta$$

so that, because of (S_6), the motion starting at $\mathbf{y}$ remains within distance of at most μ of $\mathbf{x}_e$ for all $t \geq t_0 + T$.

Since $\|\mathbf{x} - \mathbf{x}_e\| \leq r$ is compact, it is covered by definition by a finite number of $U(\mathbf{x})$, say, $U(\mathbf{x}_1), \ldots, U(\mathbf{x}_N)$. Hence define

$$T(\mu, r) = \operatorname{Max}_i \{T(\delta/2, \mathbf{x}_i)\} < \infty$$

which completes the proof.

The foregoing argument leads to a finite value of $T(\mu, r, t_0)$ if r and T are dependent on t_0. Then we can conclude: (S_2-ii) *together with* (S_4) *implies* (S_3).

These observations motivate the following definition, which is standard:

(S_7) *Uniform asymptotic stability* is equiasymptotic stability such that δ, r, T do not depend on t_0.

It is clear that uniform asymptotic stability involves the explicit assumption of uniform stability. Unlike in the case of nonuniform stability, where (S_2-ii) implies (S_3), this assumption cannot be dispensed with as the following example due to Massera shows:

Example 4. Impulse Response With Growing Peaks. Consider the first-order linear system

$$\dot{x}_1 = (4t \sin t - 2t)x_1$$

By separation of variables and integrating, the impulse response of this system (= 1×1 transition matrix) is

$$\phi_{11}(t, t_0) = \exp\,(4 \sin t - 4t \cos t - t^2 - 4 \sin t_0 + 4t_0 \cos t_0 + t_0^2)$$

which is less than

$$\exp\,[4(2 + |t| + |t_0|) - t^2 + t_0^2]$$

which shows that $\|\phi\| \to 0$ with $t \to \infty$ uniformly in t_0 (and of course uniformly in $\|\mathbf{x}_0\| \leq r$, by linearity.) However,

$$\phi_{11}((2n + 1)\pi, 2n\pi) = \exp\,[\pi(4 - \pi)(4n + 1)]$$

The "peaks" of the impulse response increase indefinitely as $t_0 \to \infty$. There is no uniform stability; in fact, the motions are not even uniformly bounded.

This example motivates also the following definition:

(S_8) An equilibrium state $\mathbf{x}_e$ of a free dynamic system is *uniformly asymptotically stable in the large* if

(i) it is uniformly stable,

(ii) it is uniformly bounded, i.e., given any $r > 0$ there is some $B(r)$ such that $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq r$ implies $\|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq B$ for all $t \geq t_0$,

(iii) every motion converges to $\mathbf{x}_e$ as $t \to \infty$ uniformly in t_0 and $\|\mathbf{x}_0\| \leq r$, when r is fixed but arbitrarily large; i.e., given any $r > 0$ and $\mu > 0$ there is some $T(\mu, r)$ such that $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq r$ implies $\|\phi(t;\ \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq \mu$ for all $t \geq 0\, t_0 + T$.

One can define a type of boundedness by taking $B(\mathbf{x}_0)$; however, this condition, with (S_8-iii), would imply (S_8-ii), by arguments similar to those concerning uniform asymptotic stability.

By a slight extension of Example 4, it follows that (S_8-iii) does not imply (S_8) in general. A case where this is true occurs when $\mathbf{f}(\mathbf{x}, t)$ satisfies a uniform Lipschitz condition, i.e.,

$$\|\mathbf{f}(\mathbf{x}, t) - \mathbf{f}(\mathbf{y}, t)\| \leq k\|\mathbf{x} - \mathbf{y}\|$$

where k is a constant independent of $\mathbf{x}$, $\mathbf{y}$, or t. To prove this assertion, one merely makes use of the relation (6).

The different types of stability, summarizing the discussion of this section, are displayed in Fig. 6.

It is apparent that concepts of stability are closely related to concepts of convergence. When there are many of the latter, there are correspondingly many types of stability. For instance, in the study of stochastic systems, one can talk about stability in probability, stability in mean, almost sure stability, etc. It is not hard in principle to apply the "second method" to such problems as has been done by Bertram and Sarachik [19].

6—The Main Theorems

The application of the "second method" to stability problems consists of defining a Lyapunov function with appropriate properties whose existence implies the desired type of stability. We shall consider uniform asymptotic stability in the large and define the class of Lyapunov functions for this case. By weakening the various requirements on Lyapunov functions, we obtain other stability theorems as a by-product. See also Massera [2], Hahn [3], and Malkin [4].

Theorem 1. (Lyapunov) *Consider the continuous-time, free dynamic system*

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, t) \tag{5-F}$$

where $\mathbf{f}(\mathbf{0}, t) = \mathbf{0}$ *for all* t.

SUPPOSE *there exists a scalar function* $V(\mathbf{x}, t)$ *with continuous first partial derivatives with respect to* $\mathbf{x}$ *and* t *such that* $V(\mathbf{0}, t) = 0$ *and*

(i) $V(\mathbf{x}, t)$ *is positive definite; i.e., there exists a continuous, nondecreasing scalar function* α *such that* $\alpha(0) = 0$ *and, for all* t *and all* $\mathbf{x} \neq \mathbf{0}$

$$0 < \alpha(\|\mathbf{x}\|) \leq V(\mathbf{x}, t)$$

(see Fig. 7);

(ii) *There exists a continuous scalar function* γ *such that* $\gamma(0) = 0$ *and the derivative* $\dot{V}$ *of* V *along the motion starting at* t, $\mathbf{x}$ *satisfies, for all* t *and all* $\mathbf{x} \neq \mathbf{0}$,

$$\left.\begin{aligned}\dot{V}(\mathbf{x}, t) &\equiv \left.\frac{dV(\mathbf{x}, t)}{dt}\right|_{\substack{\text{along motion}\\ \text{starting at } t,\ \mathbf{x}}} \\ &= \left.\frac{dV(\boldsymbol{\phi}(\tau;\ \mathbf{x}_0, t), \tau)}{d\tau}\right|_{\tau = t} \\ &= \lim_{h \to 0} [V(\mathbf{x} + h\mathbf{f}(\mathbf{x}, t), t + h) - V(\mathbf{x}, t)]/h \\ &= \partial V/\partial t + (\mathbf{grad}\ V)'\mathbf{f}(\mathbf{x}, t) \\ &\leqq -\gamma(\|\mathbf{x}\|) < 0^{6};\end{aligned}\right\} \quad (13)$$

(iii) *There exists a continuous, nondecreasing scalar function such that* $\beta(0) = 0$ *and, for all* t,

$$V(\mathbf{x}, t) \leqq \beta(\|\mathbf{x}\|)^{7};$$

(iv) $\alpha(\|\mathbf{x}\|) \to \infty$ with $\|\mathbf{x}\| \to \infty$

THEN *the equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *is uniformly asymptotically stable in the large;* $V(\mathbf{x}, t)$ *is called a Lyapunov function of the system* (5-F).

Corollary 1.1. *The following conditions are sufficient for the various weaker types of stability:*

(*a*) *Uniform asymptotic stability:* (i–iii).
(*b*) *Equiasymptotic stability in the large:* (i–ii), (iv).
(*c*) *Equiasymptotic stability:* (i–ii).
(*d*) *Uniform stability:* (i), (iii), *and* (ii$_1$): $\dot{V}(\mathbf{x}, t) \leqq 0$ *for all* $\mathbf{x}, t$.
(*e*) *Stability:* (i–ii$_1$).
(*f*) *No finite escape time:* (i), (iv), *and* (ii$_2$): $\dot{V}(\mathbf{x}, t) \leqq cV(\mathbf{x}, t)$ *for all* $\mathbf{x}, t$; c *being a positive constant.*

Corollary 1.2. *For a continuous-time autonomous dynamic system*

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}) \quad (5\text{-A})$$

$\mathbf{f}(\mathbf{0}) = \mathbf{0}$, *equiasymptotic stability in the large is assured by the existence of scalar function* $V(\mathbf{x})$ *with continuous first partial derivatives with respect to* $\mathbf{x}$, *such that* $V(\mathbf{0}) = 0$ *and*

(i-A) $V(\mathbf{x}) > 0$ *for all* $\mathbf{x} \neq \mathbf{0}$
(ii-A) $\dot{V}(\mathbf{x}) < 0$ *for all* $\mathbf{x} \neq \mathbf{0}$
(iv-A) $V(\mathbf{x}) \to \infty$ *with* $\|\mathbf{x}\| \to \infty$

Corollary 1.3. *In Corollary 1.2, Condition* (*ii-A*) *may be replaced by:*

(ii$_1$-A) $\dot{V}(\mathbf{x}) \leqq 0$ *for all* $\mathbf{x}$
(ii$_2$-A) $\dot{V}(\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0))$ *does not vanish identically in* $t \geqq t_0$ *for any* t_0 *and any* $\mathbf{x}_0 \neq \mathbf{0}$.

Proof of Theorem 1. Using (ii) we see that V is decreasing along any motion:

$$V(\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0), t) - V(\mathbf{x}_0, t_0) = \int_{t_0}^{t} \dot{V}(\boldsymbol{\phi}(\tau;\ \mathbf{x}_0, t_0), \tau)d\tau < 0, \qquad t > t_0 \quad (14)$$

(*a*) To prove uniform stability, consider any $\epsilon > 0$. Take $\delta(\epsilon) > 0$ such that $\beta(\delta) < \alpha(\epsilon)$—this is possible because β is continuous and 0 at 0; see Fig. 7. Then if $\|\mathbf{x}_0\| \leqq \delta$, t_0 being arbitrary, we have using (ii) and (14) (see Fig. 8), for all $t \geqq t_0$

$$\begin{aligned}\alpha(\epsilon) > \beta(\delta) &\geqq V(\mathbf{x}_0, t_0) \\ &\geqq V(\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0), t) \\ &\geqq \alpha(\|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\|)\end{aligned}$$

But since α is nondecreasing and positive, this implies

$$\|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\| < \epsilon;\ t \geqq t_0, \quad \|\mathbf{x}_0\| \leqq \delta$$

for arbitrary t_0, which is what was to be proved. Note that (iv) is not needed and that it is sufficient to have $\dot{V} \leqq 0$.

(*b*) Now we prove that $\|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\| \to 0$ with $t \to \infty$ uniformly in t_0 and $\|\mathbf{x}_0\| \leqq r$.

Take any positive constant c_1 and find an $r > 0$ satisfying $\beta(r) < \alpha(c_1)$. Take any initial state $\|\mathbf{x}_0\| \leqq r$. By part (*a*) of the proof, then $\|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\| \leqq c_1$ for all $t \geqq t_0$, t_0 being arbitrary.

Now take any $0 < \mu \leqq \|\mathbf{x}_0\|$. Find a $\nu(\mu) > 0$ such that $\beta(\nu) < \alpha(\mu)$. Denote by $c_2(\mu, r) > 0$ the minimum of the continuous function $\gamma(\|\mathbf{x}\|)$ on the compact set $\nu(\mu) \leqq \|\mathbf{x}\| \leqq c_1(r)$. Define $T(\mu, r) = \beta(r)/c_2(\mu, r) > 0$.

Suppose now that $\|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\| > \nu$ over the interval $t_0 \leqq t \leqq t_1 = t_0 + T$. Again by (ii) and (14), we have

$$\begin{aligned}0 < \alpha(\nu) &\leqq V(\boldsymbol{\phi}(t_1;\ \mathbf{x}_0, t_0), t_1) \\ &\leqq V(\mathbf{x}_0, t_0) - (t_1 - t_0)c_2 \\ &\leqq \beta(r) - Tc_2 = 0\end{aligned}$$

which is a contradiction. Hence for some t in the interval t_0, t_1, say t_2, we have $\|\mathbf{x}_2\| = \|\boldsymbol{\phi}(t;\ \mathbf{x}_0, t_0)\| = \nu$. Therefore

$$\begin{aligned}\alpha(\|\boldsymbol{\phi}(t;\ \mathbf{x}_2, t_2)\|) &\leqq V(\boldsymbol{\phi}(t;\ \mathbf{x}_2, t_2), t) \\ &\leqq V(\mathbf{x}_2, t_2) \\ &\leqq \beta(\nu) < \alpha(\mu)\end{aligned}$$

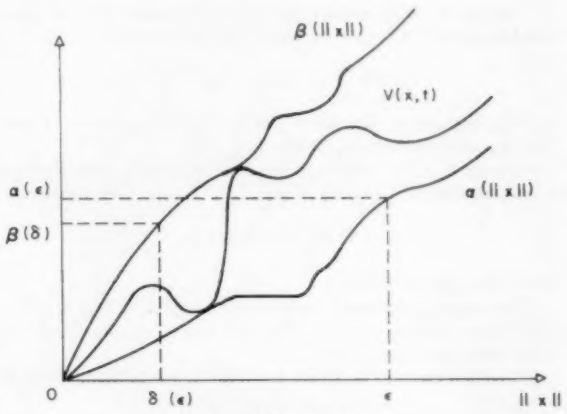
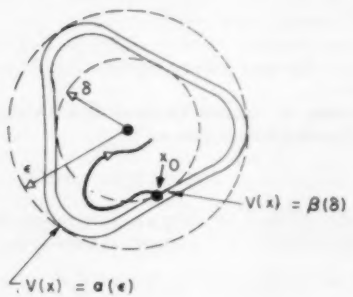


Fig. 7 Definition of V, α, β



$$\alpha(\|\mathbf{x}(t)\|) \leqq V(\mathbf{x}(t)) \leqq V(\mathbf{x}_0) \leqq \beta(\delta) < \alpha(\epsilon)$$

Fig. 8 Proof of asymptotic stability

[6] This is slightly weaker than the usual requirement that $\dot{V}$ be negative definite, since we do not need the hypothesis that γ is nondecreasing.

[7] In the literature, particularly the Russian, this condition is sometimes called: "V has an infinitely small upper bound."

for all $t \geq t_0$. Hence

$$\|\phi(t;\ \mathbf{x}_0,\ t_0)\| < \mu \quad \text{for all } t \geq t_0 + T(\mu, r) \geq t_0$$

which proves uniform asymptotic stability.

(c) To prove uniform asymptotic stability in the large, i.e., that the constant r can be chosen to be arbitrarily large, observe that by (iv) there exists for any r a constant $c_1(r)$ such that $\beta(r) < \alpha(c_1)$.

Moreover, uniform boundedness is automatic, on setting $B(r) = c_1(r)$. Q.E.D.

Proof of Corollary 1.1. (a) This is contained in the proof of Theorem 1.

(b) Let $\beta^*(\|\mathbf{x}\|, t)$ be the maximum of $V(\mathbf{y}, t)$ for $\|\mathbf{y}\| \leq \|\mathbf{x}\|$. Replacing $\beta(\|\mathbf{x}\|)$ by $\beta^*(\|\mathbf{x}\|, t)$, the proof of Theorem 1 proceeds as before, but now δ and T are also functions of t_0.

(c) Use β^* instead of β and omit Part (c) of the proof of Theorem 1.

(d) Part (a) of the proof of Theorem 1.

(e) Use β^* defined previously and proceed as in Part (a) of the proof of Theorem 1. Then δ is also a function of t_0.

(f) An easy calculation shows that (ii_3) implies

$$V(\phi(t;\ \mathbf{x}_0,\ t_0)) \leq e^{c(t-t_0)} V(\mathbf{x}_0,\ t_0)$$

and by (i)

$$\alpha(\|\phi(t;\ \mathbf{x}_0,\ t_0)\|) \leq e^{c(t-t_0)} V(\mathbf{x}_0,\ t_0) < \infty$$

Because of (iv), $\|\mathbf{x}\| = \infty$ implies $\alpha(\|\mathbf{x}\|) = \infty$ which is impossible for any finite t. Q.E.D.

Proof of Corollary 1.2. By (iv), given any $a > 0$ there is a $b(a) > 0$ such that $V(\mathbf{x}) > a$ when $\|\mathbf{x}\| > b$. (If (iv-A) is missing and only local stability is of interest, pick $b > 0$ arbitrarily.) Let

$$c(\|\mathbf{x}\|) = \{\operatorname{Min}_{\mathbf{y}} V(\mathbf{y});\ \|\mathbf{y}\| = \|\mathbf{x}\|\}$$

and then define

$$\alpha(\|\mathbf{x}\|) = \operatorname{Min}_{\mathbf{y}} \{V(\mathbf{y});\ \|\mathbf{x}\| \leq \|\mathbf{y}\| \leq b(c(\|\mathbf{x}\|))\}$$

$$\beta(\|\mathbf{x}\|) = \operatorname{Max}_{\mathbf{y}} \{V(\mathbf{y});\ \|\mathbf{y}\| \leq \|\mathbf{x}\|\}$$

$$\gamma(\|\mathbf{x}\|) = \operatorname{Min}_{\mathbf{y}} \{-\dot{V}(\mathbf{y});\ \|\mathbf{y}\| = \|\mathbf{x}\|\}$$

The functions α, β, γ obviously satisfy the hypotheses of Theorem 1. Q.E.D.

Proof of Corollary 1.3. It suffices to show that (ii_1-A), (ii_2-A), together with (i-A), (iv-A) imply asymptotic stability in the large for stationary systems. By (iv-A) and (ii_1-A), for every $\mathbf{x}_0$ there is a constant $c_1(\|\mathbf{x}_0\|)$ such that $\|\phi(t;\ \mathbf{x}_0,\ t_0)\| \leq c_1$ for all $t \geq t_0$, t_0 being arbitrary since the system is stationary.

Suppose $\|\phi(t;\ \mathbf{x}_0,\ t_0)\| \geq \nu > 0$. By (i-A) and ($\text{ii}_1$-A) the integral

$$-V(\mathbf{x}_0) < \int_{t_0}^{t} \dot{V}(\phi(\tau;\ \mathbf{x}_0,\ t_0))d\tau \leq 0$$

is nonincreasing and bounded as $t \to \infty$, which implies that $\dot{V}$ is identically zero on the positive limit set [11] of $\mathbf{x}_0$, which is contained in the region $\nu \leq \|\mathbf{x}\| \leq c_1$. Since the limit set is itself a motion, this contradicts (ii_2-A). Hence every motion will eventually enter the set $\|\mathbf{x}\| \leq \nu$. But if ν is chosen so that $\beta(\nu) < \alpha(\mu)$, then asymptotic stability in the large follows, as in the proof of Theorem 1. Q.E.D.

As already pointed out, there is no explicit prescription in general for finding Lyapunov functions satisfying the requirements of Theorem 1 or of its corollaries. For this reason, it is very important to know when a Lyapunov function with certain properties exists, so that the search for one should not be in vain! It turns out that this question is closely related to the concept of the uniform asymptotic stability. We now quote what is at present the best available result along these lines [2]:

Theorem 2. (Massera) *Let the function* $\mathbf{f}$ *defining the differential equation* (5-F) *be Lipschitzian. Assume also that* $\mathbf{f}(\mathbf{0}, t) = \mathbf{0}$ *and that the equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *is uniformly asymptotically stable in the large.*

Then there exists a Lyapunov function $V(\mathbf{x}, t)$ *which is infinitely differentiable with respect to* $\mathbf{x}$, t, *and satisfies all of the hypotheses of Theorem 1.*

Thus the existence of a Lyapunov function as required by Theorem 1 is necessary and sufficient for uniform asymptotic stability in the large.

PROOF: See Reference [2], Theorem 23*.

The stability problem in nonstationary linear systems is one of considerable subtlety. Roughly speaking, the requirement of "uniformity" must be present in the formal definitions of stability, in order that it should coincide with the intuitive concept of stability. The troubles are illustrated by the following example:

Example 5. Stable System With Unbounded Step Response. Consider the nonstationary linear system

$$\dot{x}_1 + x_1/t = y_1(t)$$

for values of $0 < t_0 \leq t$. (For values of t_0 below 0, the equation has finite escape time (see Example 3), and therefore does not define a dynamic system.) The impulse response (which in this case is also the 1×1 transition matrix) of this system is

$$\phi_{11}(t, t_0) = t_0/t;\ t \geq t_0 > 0$$

The system is evidently asymptotically stable, but not uniformly asymptotically stable: If $t_0 > 0$ and $T = (t - t_0)$ is a fixed constant then

$$\phi_{11}(t, t_0) = t_0/(T + t_0) > 1/2$$

for t_0 sufficiently large, no matter how large T is.

Now suppose that the initial state at some $t_0 > 0$ is zero. Let the input $y_1(t)$ be a unit step function occurring at $t = t_0$. Then the unit step response is given by

$$x_1(t) = \int_{t_0}^{t} \phi_{11}(t, \tau)d\tau$$

$$= (t - t_0^2/t)/2 \qquad (t \geq t_0 > 0)$$

which tends to infinity with t!

The difficulty here is due to the fact that the impulse response does not tend to 0 exponentially as in an asymptotically stable linear *stationary* system. It so happens that, in the linear case, exponential convergence to equilibrium is equivalent to uniform asymptotic stability. The latter, in turn, is equivalent to the existence of a Lyapunov function or the fact that every bounded excitation produces a bounded response. These facts are summarized by:

Theorem 3. (Lyapunov-Perron-Malkin) *Consider a continuous-time, linear dynamic system*

$$d\mathbf{x}/dt = \mathbf{F}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}(t) \qquad (5\text{-L})$$

subject to the restrictions

(i) $\|\mathbf{F}(t)\| \leq c_1$[8] $< \infty$ *for all* t

(ii) $0 < c_2 \leq \|\mathbf{D}(t)\mathbf{x}\| \leq c_3 < \infty$ *for all* $\|\mathbf{x}\| = 1$, *all* t

THEN *the following propositions concerning this system are equivalent:*

(A) *Any uniformly bounded excitation*

$$\|\mathbf{u}(t)\| \leq c_4 < \infty \quad (t \geq t_0)$$

[8] In this theorem and its proof, $c_1, c_2, \ldots$ are used to denote various fixed positive constants. We continue to use this convention in the sequel.

gives rise to a uniformly bounded response for all $t \geq t_0$

$$\|\mathbf{x}(t)\| = \left\|\Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{D}(\tau)\mathbf{u}(\tau)d\tau\right\| \leq c_3(c_4, \|\mathbf{x}_0\|) < \infty; \tag{14-A}$$

(*B*) *For all* $t \geq t_0$, $\int_{t_0}^{t} \|\Phi(t, \tau)\|d\tau \leq c_4 < \infty$;

(*C*) *The equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *of the free system is uniformly asymptotically stable;*

(*D*) *There are positive constants* c_7, c_8 *such that, whenever* $t \geq t_0$,

$$\|\Phi(t, t_0)\| \leq c_7 e^{-c_8(t-t_0)}; \tag{15}$$

(*E*) *Given any positive-definite matrix* $\mathbf{Q}(t)$ *continuous in* t *and satisfying for all* $t \geq t_0$

$$0 < c_9\mathbf{I} \leq \mathbf{Q}(t) \leq c_{10}\mathbf{I} < \infty^{9} \tag{16}$$

the scalar function defined by

$$V(\mathbf{x}, t) = \int_{t}^{\infty} \|\Phi(\tau, t)\mathbf{x}\|^2_{\mathbf{Q}(\tau)}d\tau^{10} \tag{17}$$

$$= \|\mathbf{x}\|^2_{\mathbf{P}(t)} \tag{18}$$

exists[11] *and is a Lyapunov function for the free system satisfying the requirements of Theorem 1, with its derivative along the free motion starting at* $\mathbf{x}$, t *being*

$$\dot{V}(\mathbf{x}, t) = -\|\mathbf{x}\|^2_{\mathbf{Q}(t)}$$

Propositions (*A*)-(*D*) relate to different abstract notions of stability. The equivalence of (*A*) and (*B*) was first proved by Perron [20]. Zadeh used (*A*) as his basic definition of stability [21]. By proposition (*E*), the preceding notions of stability are equivalent to the existence of a Lyapunov function. Malkin [22] was the first to prove that (*C*) is equivalent to (*E*); his method of proof implied condition (*B*), hence, by Perron's result, the proof of the entire theorem. See also Antosiewicz and Davis [23]. The proof given below, based on unpublished work of Kalman. is much simpler than the original proof of Malkin and Perron, due to the systematic use of the transition matrix.

None of the propositions in Theorem 3 offers a method of determining whether a given system is uniformly asymptotically stable or not without computing its transition matrix for all t, t_0. Such a method is known so far only in the case of linear stationary systems and is already contained in Lyapunov's original memoire:

Corollary 3.1. (Lyapunov) *The equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *of a continuous-time, free, linear, stationary dynamic system*

$$d\mathbf{x}/dt = \mathbf{F}\mathbf{x} \tag{5-FLS}$$

is asymptotically stable (*a*) *if and* (*b*) *only if given any symmetric, positive-definite matrix* $\mathbf{Q}$ *there exists a symmetric, positive-definite matrix* $\mathbf{P}$ *which is the unique solution of the set of* $n(n+1)/2$[12] *linear equations*

$$\mathbf{F}'\mathbf{P} + \mathbf{P}\mathbf{F} = -\mathbf{Q} \tag{19}$$

and $\|\mathbf{x}\|^2_{\mathbf{P}}$ *is a Lyapunov function for* (5-FLS).

A slight extension of this result is useful for some purposes:

Corollary 3.2. (Kalman) *The real parts of the eigenvalues of a (constant) matrix* $\mathbf{F}$ *are* $< \sigma$ *if and only if given any symmetric, positive-definite matrix* $\mathbf{Q}$ *there exists a symmetric, positive-definite matrix* $\mathbf{P}$ *which is the unique solution of the set of* $n(n+1)/2$ *linear equations.*

$$-2\sigma\mathbf{P} + \mathbf{F}'\mathbf{P} + \mathbf{P}\mathbf{F} = -\mathbf{Q} \tag{20}$$

Proof of Theorem 3. (Kalman) The proof consists mainly of using the multiplicative property (8-L) of transition matrices and various technicalities involving norms and positive-definite matrices.

To show that (*A*) implies (*B*). If (*B*) did not hold, for at least one pair (i^*, j^*) of subscripts the integral

$$\int_{t_0}^{t} |\phi_{i^*j^*}(t, \tau)|d\tau$$

must tend to ∞ as $(t - t_0) \to \infty$. By (ii), $1/\epsilon_3 \leq \|\mathbf{D}^{-1}(\tau)\| \leq 1/c_2$ for all τ. Therefore define, for all $t_0 \leq \tau \leq t$

$$u_k(\tau) = |\mathbf{D}^{-1}(\tau)|_{k j^*} \operatorname{sgn} \phi_{i^*j^*}(t, \tau)$$

where

$$\operatorname{sgn} x = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{21}$$

Then $u(\tau)$ is bounded; we can take $c_4 = 1/c_2$. Using this and (75), with $\mathbf{x}_0 = \mathbf{0}$,

$$\begin{aligned} \infty > c_3(c_4, 0) &\geq \|\mathbf{x}(t)\| \\ &\geq n^{-1/2} \sum_{i=1}^{n} |x_i(t)| \\ &\geq \left|\sum_{k=1}^{n} \int_{t_0}^{t} \phi_{i^*j^*}(t, \tau)d_{j^*k}(\tau)u_k(\tau)d\tau\right| \\ &= \int_{t_0}^{t} |\phi_{i^*j^*}(t, \tau)|d\tau \end{aligned}$$

By assumption, the last integral tends to infinity with $(t - t_0)$, which is a contradiction.

To show that (*B*) implies (*C*). (This is the only nontrivial part of the proof.) First we show that (*B*) implies

$$\|\Phi(t, t_0)\| \leq c_{11} < \infty \qquad (t \geq t_0) \tag{22}$$

By (i) and (*B*) we can write

$$\begin{aligned} \infty > c_1c_6 &\geq \int_{t_0}^{t} \|\Phi(t, \tau)\| \cdot \|\mathbf{F}(\tau)\|d\tau \\ &\geq \left\|\int_{t_0}^{t} \Phi(t, \tau)\mathbf{F}(\tau)d\tau\right\| \end{aligned}$$

Since the transition matrix satisfies its own differential equation (10), we obtain, after using repeatedly (8-L)

$$-\frac{d}{d\tau}\Phi^{-1}(\tau, t) = \Phi^{-1}(\tau, t)\mathbf{F}(\tau) = \Phi(t, \tau)\mathbf{F}(\tau)$$

Therefore

$$\begin{aligned} \infty > c_1c_8 &\geq \left\|\int_{t_0}^{t} \left[\frac{d\Phi(t, \tau)}{d\tau}\right]d\tau\right\| \\ &= \|\Phi(t, t_0) - \mathbf{I}\|, \qquad t \geq t_0 \end{aligned}$$

which implies (22).

Using (*B*), (22), and (8-L),

$$\begin{aligned} \infty > c_6c_{11} &\geq \int_{t_0}^{t} \|\Phi(t, \tau)\| \cdot \|\Phi(\tau, t_0)\|d\tau \\ &\geq \|\Phi(t, t_0)\|(t - t_0), \qquad t \geq t_0 \end{aligned}$$

[9] If $\mathbf{A}$, $\mathbf{B}$ are two matrices, we write $\mathbf{A} < \mathbf{B}$[$\mathbf{A} \leq \mathbf{B}$] to express the fact that $\mathbf{B} - \mathbf{A}$ is positive-definite [semidefinite].

[10] We write (see Appendix) $\mathbf{x}'\mathbf{A}\mathbf{x} = \|\mathbf{x}\|^2_{\mathbf{A}}$ where $\mathbf{A}$ is a symmetric, positive-definite matrix.

[11] I.e., the integral defined by (17) is finite for all finite values of $\mathbf{x}$ and t.

[12] Since $\mathbf{P}$ is symmetric, it has precisely this many unknown elements in (19).

Letting

$$T(\mu, r) = \frac{c_6 c_{11}}{\mu} r,$$

it follows that

$$\|\Phi(t, t_0)\mathbf{x}_0\| \leq \mu$$

when $\|\mathbf{x}_0\| \leq r$ and $t \geq T + t_0$, so that we have uniform asymptotic stability.

To show that (C) implies (D). Since uniform asymptotic stability implies uniform stability (see definitions (S_7), (S_8)), the transition matrix is uniformly bounded. In fact, take any $\epsilon > 0$, $0 < \|\mathbf{x}_0\| \leq \delta(\epsilon)$. By linearity:

$$\delta^{-1}\|\phi(t;\ \mathbf{x}_0, t_0)\| \leq \|\Phi(t, t_0)\| \cdot \delta^{-1}\|\mathbf{x}_0\| \leq \|\Phi(t, t_0)\| \leq \delta^{-1}\epsilon = c_7/2$$

By uniform asymptotic stability, let $T = T(2^{-1}, 1)$ so that

$$\|\Phi(t_0 + T, t_0)\| \leq 1/2$$

independently of t_0. By induction and (77)

$$\|\Phi(t_0 + kT, t_0)\| \leq \|\Phi(t_0 + kT, t_0 + (k-1)T)\| \ldots \|\Phi(t_0 + T, t_0)\| \leq 2^{-k}$$

Hence if $c_8 = (\log 2)/T > 0$. then

$$\|\Phi(t, t_0)\| \leq c_7 e^{-c_8(t - t_0)} \qquad (t \geq t_0) \tag{15}$$

To show that (D) implies (A). By (15), (ii), and (78)

$$\left\|\phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{D}(\tau)\mathbf{u}(\tau)d\tau\right\| \leq c_{11}\|\mathbf{x}_0\| + c_2 c_4 c_7 \int_{t_0}^{t} e^{-c_8(t-\tau)}\, d\tau \leq c_{11}\|\mathbf{x}_0\| + c_2 c_4 c_7 c_8^{-1} < \infty$$

so that (14-A) holds.

By Theorem 1, (E) implies (C).

To show that (D) implies (E). This is essentially Theorem 2; we give an independent proof. Proceeding as above and using (16)

$$V(\mathbf{x}, t) \leq c_{10} \int_t^\infty \|\Phi(\tau, t)\mathbf{x}\|^2 d\tau$$

By (15)

$$\leq \frac{c_7^2 c_{10}}{2c_8}\|\mathbf{x}\|^2 = \beta(\|\mathbf{x}\|) < \infty \tag{23}$$

Further, using (i), (16), and (10)

$$V(\mathbf{x}, t) \geq c_1^{-2} c_9 \int_t^\infty \|\mathbf{F}(\tau)\|^2 \|\Phi(\tau, t)\mathbf{x}\|^2 d\tau \geq c_1^{-2} c_9 \|\mathbf{x}\|^2 = \alpha(\|\mathbf{x}\|) > 0, \qquad \mathbf{x} \neq \mathbf{0}$$

The derivative $\dot{V}$ of the Lyapunov function V along the free motion starting at t, $\mathbf{x}$ is obtained by direct computation

$$\left.\begin{aligned} \dot{V}(\mathbf{x}, t) &= \lim_{h \to 0} \frac{1}{h}\left[V(\phi(t + h;\ \mathbf{x}, t),\ t + h) - V(\mathbf{x}, t)\right] \\ &= \lim_{h \to 0} -\frac{1}{h}\left[\int_t^{t+h} \|\Phi(\tau, t)\mathbf{x}\|^2_{\mathbf{Q}(\tau)} d\tau\right] \\ &= -\|\mathbf{x}\|^2_{\mathbf{Q}(t)} \leq -c_9\|\mathbf{x}\|^2 = -\gamma(\|\mathbf{x}\|) < 0 \quad \text{when} \quad \mathbf{x} \neq \mathbf{0} \end{aligned}\right\} \tag{24}$$

This completes the proof of Theorem 3.

Proof of Corollary 3.1. (a) Let $\mathbf{P} > \mathbf{0}$ satisfy (19). Define

$$V(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{P}} = \mathbf{x}'\mathbf{P}\mathbf{x} > 0 \qquad \text{when } \mathbf{x} \neq \mathbf{0}$$

as a tentative Lyapunov function for system (5-FLS). Then, by (13)

$$\dot{V}(\mathbf{x}) = (\text{grad } V)'\mathbf{F}\mathbf{x} = 2(\mathbf{P}\mathbf{x})'\mathbf{F}\mathbf{x} = \mathbf{x}'(\mathbf{F}'\mathbf{P} + \mathbf{P}\mathbf{F})\mathbf{x} \tag{25}$$

and by (19)

$$\dot{V}(\mathbf{x}) = -\mathbf{x}'\mathbf{Q}\mathbf{x} = -\|\mathbf{x}\|^2_{\mathbf{Q}} < 0 \text{ when } \mathbf{x} \neq \mathbf{0}$$

so that V satisfies the hypotheses of Corollary 1.2; therefore (5-FLS) is asymptotically stable.

(b) Suppose (5-FLS) is asymptotically stable. Since it is stationary, it is then also uniformly asymptotically stable. Therefore, by Theorem 3 (C) and (E), the matrix $\mathbf{P}(t)$ defined by (18) exists. Assuming $\mathbf{Q}(t) = \mathbf{Q} = \text{const}$ it follows then by the stationarity of (5-FLS) that $\mathbf{P}(t) = \mathbf{P} = \text{const}$. Also, the derivative of the Lyapunov function $V(\mathbf{x})$ defined by (18) is $-\mathbf{x}'\mathbf{Q}\mathbf{x}$, which is equal to (25), so that $\mathbf{P}$ satisfies (19).

To show that $\mathbf{P}$ defined by (18) is the unique solution of (19) let $\bar{\mathbf{P}}$ be any solution of (19). Then, by the definition (11) of the transition matrix in the stationary case

$$\mathbf{P} = \int_0^\infty (\exp \mathbf{F}'\tau)\mathbf{Q}(\exp \mathbf{F}\tau)d\tau = -\int_0^\infty (\exp \mathbf{F}'\tau)(\mathbf{F}'\bar{\mathbf{P}} + \bar{\mathbf{P}}\mathbf{F})(\exp \mathbf{F}\tau)d\tau$$

Since $\mathbf{F}$ commutes with $\exp \mathbf{F}\tau$ for all τ

$$\mathbf{P} = -\int_0^\infty \frac{d}{d\tau}[(\exp \mathbf{F}'\tau)\bar{\mathbf{P}}(\exp \mathbf{F}\tau)]d\tau = -(\exp \mathbf{F}'\tau)\bar{\mathbf{P}}(\exp \mathbf{F}\tau)\Big|_0^\infty = \bar{\mathbf{P}}$$

the last equality following from asymptotic stability. Q.E.D.

Incidentally, Bellman [24] proves that (19) has a unique (but not necessarily positive definite) solution for any $\mathbf{Q}$ whenever $\lambda_i(\mathbf{F}) + \lambda_j(\mathbf{F}) \neq 0$ for all i, j.

Proof of Corollary 3.2. The assumptions on $\mathbf{F}$ require that

$$Re\lambda_i(\mathbf{F}) < \sigma$$

From the characteristic equation of a matrix it follows that this is equivalent to:

$$Re\lambda_i(\mathbf{F} - \sigma\mathbf{I}) < 0$$

Replacing $\mathbf{F}$ in (19) by $\mathbf{F} - \sigma\mathbf{I}$ yields (20). (Notice that this argument is equivalent to the "shift" theorem in the theory of the Laplace transform.) Q.E.D.

7—Applications to Stability Theory

7.1 Routh-Hurwitz Conditions for Linear Stationary Systems. Analogously to the various methods associated with the names of Routh and Hurwitz, Corollary 3.1 provides a purely algebraic procedure [25] for testing whether a given free, linear, stationary system is asymptotically stable:

(a) Take an arbitrary symmetric, positive-definite matrix $\mathbf{Q}$ (say, the unit matrix);

(b) Solve (19) for the unknown components of the symmetric matrix $\mathbf{P}$. This is a set of $n(n+1)/2$ linear equations in the unknown elements $p_{11}, \ldots, p_{1n}; p_{22}, \ldots, p_{2n}; \ldots; p_{nn}$ of $\mathbf{P}$. (The requirement that this set of equations have a nonzero determinant is implied by the Routh-Hurwitz inequalities.)

(c) Check to see if $\mathbf{P}$ is positive-definite. It suffices to check if all leading principal minors[18] of $\mathbf{P}$ are positive. This supplies the Routh-Hurwitz inequalities.

[18] I.e., the determinants of the following submatrices of $\mathbf{P}$: (1) $\mathbf{P}$ itself, (2) $\mathbf{P}$ with the nth row and column deleted, . . ., (n) the 1-1 matrix consisting of p_{11}.

This procedure can be used to derive the Routh-Hurwitz inequalities as in Example 6. However, this is mainly of theoretical interest since the derivations are very complicated when the matrix $\mathbf{F}$ contains more than about 4 or 5 nonzero coefficients.

The principal significance of this procedure is that it provides a convenient and efficient algorithm to determine stability by machine computation. This is far from trivial in the case of large-scale systems since the Routh-Hurwitz conditions in their usual form apply only to a single, nth-order equation (which has in general n different coefficients). Before the Routh-Hurwitz conditions can be applied to a general linear system of n first-order equations (which in general has n^2 different coefficients), it is necessary to compute the characteristic equation of the matrix $\mathbf{F}$; this is far from a trivial job when the number of nonzero coefficients in $\mathbf{F}$ is large. These difficulties are avoided when (19) is used. Another efficient computation procedure for determining stability of general linear stationary systems, due to Schwarz [26], reduces $\mathbf{F}$ to a *canonic form* by means of elementary matrix transformations. This is also related to the "second method" as Example 7 shows.

The concept of a Lyapunov function as in Corollary 3.1 provides a convenient link between the Routh-Hurwitz conditions and other aspects of control-system theory, such as the integral of the quadratic error. This connection, which has come to surface only recently [27], answers an oft-heard but superficial dictum in elementary courses or textbooks, that "the Routh-Hurwitz criteria do not provide a *quantitative* measure of stability." Much further research is needed to completely elucidate these connections.

Example 6. Routh-Hurwitz Conditions for General Second-Order Case. Let $\mathbf{Q} = \mathbf{I}$ and write out explicitly (19) in the case when $\mathbf{F}$ is a 2×2 matrix:

$$\begin{bmatrix} f_{11} & f_{21} \\ f_{12} & f_{22} \end{bmatrix} \cdot \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \cdot \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

This system of equations is equivalent to

$$\begin{bmatrix} 2f_{11} & 2f_{21} & 0 \\ f_{12} & f_{11}+f_{22} & f_{21} \\ 0 & 2f_{12} & 2f_{22} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{22} \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}$$

which has a unique solution if and only if the determinant of the matrix $\mathfrak{F}$ on the left is nonzero:

$$\begin{aligned} \det \mathfrak{F} &= 4(f_{11}+f_{22})(f_{11}f_{22}-f_{12}f_{21}) \\ &= n^3(\operatorname{tr}\mathbf{F})(\det\mathbf{F}) \neq 0 \end{aligned} \tag{26}$$

Assuming (26), after tedious elementary computations we find

$$\mathbf{P} = \frac{-1}{2(\operatorname{tr}\mathbf{F})(\det\mathbf{F})} \begin{bmatrix} \det\mathbf{F} + f_{21}^2 + f_{22}^2 & -(f_{12}f_{22}+f_{21}f_{11}) \\ -(f_{12}f_{22}+f_{21}f_{11}) & \det\mathbf{F} + f_{11}^2 + f_{12}^2 \end{bmatrix} \tag{27}$$

This matrix is positive definite if and only if

$$\det\mathbf{P} = \frac{(f_{11}+f_{22})^2 + (f_{12}-f_{21})^2}{2(\operatorname{tr}\mathbf{F})^2(\det\mathbf{F})} > 0 \tag{28}$$

$$p_{11} = -\frac{\det\mathbf{F} + f_{21}^2 + f_{22}^2}{2(\operatorname{tr}\mathbf{F})(\det\mathbf{F})} > 0 \tag{29}$$

But (28) implies that

$$\det\mathbf{F} = f_{11}f_{22} - f_{12}f_{21} > 0 \tag{30}$$

and this fact, with (29), implies that

$$\operatorname{tr}\mathbf{F} = f_{11} + f_{22} < 0 \tag{31}$$

and (30–31) together imply (26).

To put these results in a better-known form, let

$$\mathbf{F} = \begin{bmatrix} 0 & 1 \\ -a_1 & -a_2 \end{bmatrix} \tag{32}$$

Then (30) and (31) require that $a_1 > 0$, $a_2 > 0$ which are the well-known Routh-Hurwitz inequalities for a constant-coefficient linear differential equation of second order.

Example 7. Canonic Form for Linear Stationary Systems. Consider the system shown in Fig. 9. The infinitesimal state-transition matrix $\mathbf{F}$ of this system is:

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ -a_1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & -a_2 & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & -a_{n-1} & -a_n \end{bmatrix} \tag{33}$$

In a little-known paper, Schwarz [26] has proved the following: *If a real constant matrix* $\mathbf{F}$ *is similar to a matrix of form* (33) (i.e., $\mathbf{F} = \mathbf{T}^{-1}\mathbf{C}\mathbf{T}$) *then the number of eigenvalues* $\lambda_i(\mathbf{F})$ *of* $\mathbf{F}$ *which have negative real parts is equal to the number of positive terms in the sequence*

$$a_n, \quad a_n a_{n-1}, \quad \dots, \quad a_n a_{n-1} \dots a_1$$

provided none of the a_i *is zero.*

Under mild restrictions Schwarz [26] also gives an algorithm for computing $\mathbf{C}(\mathbf{F})$ by means of elementary transformations.

When all the a_i are positive, it is easy to show that the system

$$d\mathbf{x}/dt = \mathbf{C}\mathbf{x} \tag{34}$$

is asymptotically stable. (The matrix (32) is a special case of this!) In fact, let $\mathbf{P}$ be the matrix whose elements on the main diagonal are given

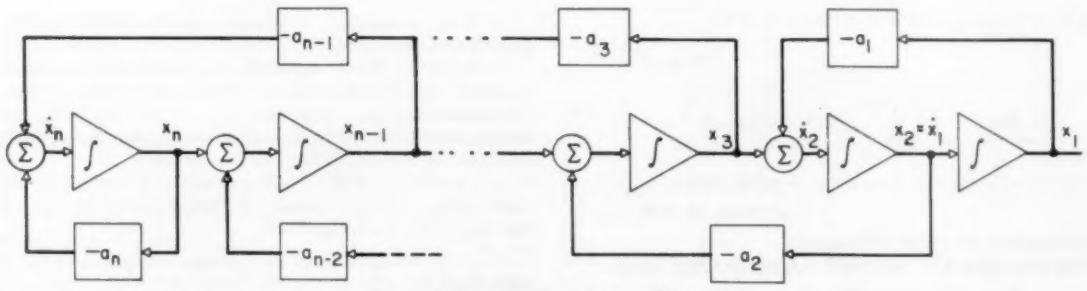


Fig. 9 System corresponding to matrix (33)

$$p_{ii} = a_n a_{n-1} \dots a_i \qquad (i = 1, \dots, n)$$

while all other elements are zero. If all the a_i are positive, then $\mathbf{P}$ is clearly positive definite. Let $V(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{P}}$. This function clearly satisfies hypotheses (i-A) and (iv-A) of Corollary 1.2. Further, a simple computation shows:

$$\dot{V}(\mathbf{x}) = \dot{\mathbf{x}}'(\mathbf{grad}\ V) = 2\mathbf{x}'\mathbf{C}'\mathbf{P}\mathbf{x} = -2a_n x_n^2 \leqq 0$$

which satisfies hypothesis (ii_1-A) of Corollary 1.3. To verify hypothesis (ii_2-A), it is necessary to show that $\dot{V}(\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0)$ is not identically zero in t for any initial state $\mathbf{x}_0 \neq \mathbf{0}$. Now if $\phi_n^2(t;\ \mathbf{x}_0,\ t_0)$ is to be identically zero in t, then by inspection of Fig. 9 it follows that $d\phi_n(t;\ \mathbf{x}_0,\ t_0)/dt$ must be identically zero in t. This in turn requires that $\phi_{n-1}(t;\ \mathbf{x}_0,\ t_0)$ be identically zero in t. Proceeding in this fashion, we see that every component of $\boldsymbol{\phi}(t;\ \mathbf{x}_0,\ t_0)$ must vanish identically in t, i.e., $\mathbf{x}_0$ is an equilibrium state of (34). Because

$$\det \mathbf{C} = \begin{cases} -\ a_1 a_3 \dots a_n & n = \text{odd} \\ +\ a_1 a_3 \dots a_{n-1} & n = \text{even} \end{cases} \neq 0$$

the system (34) has but one equilibrium state which is $\mathbf{x}_e = \mathbf{0}$. Thus $\dot{V}$ does not vanish identically along any motion starting at $\mathbf{x}_0 \neq \mathbf{0}$.

7.2 Stability of Linear Nonstationary Systems. No procedure similar to Corollary 3.1 is available so far which would permit determination of the stability of general linear nonstationary systems in an algebraic way. There is little hope that this state of affairs will change soon. It is therefore necessary to use ingenuity in finding explicit Lyapunov functions. Without discussing this matter in detail (see References [28], [29]) let us outline briefly an obvious procedure.

Take a matrix $\mathbf{P}(t)$ which is a differentiable function of t subject to

$$0 < c_1\mathbf{I} \leqq \mathbf{P}(t) \leqq c_2\mathbf{I} < \infty$$

for all t. Then the function

$$V(\mathbf{x}, t) = \|\mathbf{x}\|^2_{\mathbf{P}(t)} \tag{35}$$

obviously satisfies hypotheses (i), (iii), (iv) of Theorem 1 (in fact, $\alpha(\|\mathbf{x}\|) = c_1\|\mathbf{x}\|^2$, $\beta(\|\mathbf{x}\|) = c_2\|\mathbf{x}\|^2$). If, in addition,

$$-\mathbf{Q}(t) = \mathbf{F}'(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}(t) + \dot{\mathbf{P}}(t) \leqq -c_3\mathbf{I} < 0 \tag{36}$$

then $$\dot{V}(\mathbf{x}, t) = -\|\mathbf{x}\|^2_{\mathbf{Q}(t)}$$

satisfies hypothesis (ii) of Theorem 1, and V defined by (35) is a Lyapunov function.

The problem, of course, is to choose $\mathbf{P}(t)$ in such a way that $\mathbf{Q}(t)$ is positive definite for the largest possible region of parameters $\mathbf{F}(t)$. At first, it is best to fix $\mathbf{Q}$ and use Corollary 3.1 to get an initial guess for $\mathbf{P}$.

Example 8. Time-Varying Network. Consider the time-varying network shown in Fig. 10. The differential equations governing the dynamics are:

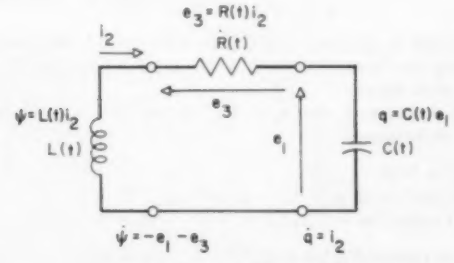


Fig. 10 Time-varying network

$$\left.\begin{aligned} \dot{x}_1 &= L^{-1}(t)x_2 \\ \dot{x}_2 &= -C^{-1}(t)x_1 - R(t)L^{-1}(t)x_2 \end{aligned}\right\} \tag{37}$$

where $x_1 = q$ = charge on capacitor and $x_2 = \psi$ = flux in inductor. Let

$$\mathbf{P}(t) = \begin{bmatrix} R(t) + 2L(t)/R(t)C(t) & 1 \\ 1 & 2/R(t) \end{bmatrix}$$

and it follows immediately from (36) that

$$\mathbf{Q}(t) = \begin{bmatrix} 2C^{-1}(t) & 0 \\ 0 & 2L^{-1}(t) \end{bmatrix} - \dot{P}(t)$$

Assuming

$$\left.\begin{aligned} 0 < \epsilon_1 \leqq L(t) \leqq \nu_1 < \infty \\ 0 < \epsilon_2 \leqq C(t) \leqq \nu_2 < \infty \\ 0 < \epsilon_3 \leqq R(t) \leqq \nu_3 < \infty \end{aligned}\right\} t \geqq t_0$$

the system (37) is asymptotically stable if for all $t \geqq t_0$

$$0 < \epsilon_4 \leqq 1 + \dot{R}(L/R^2 - C/2) + \dot{C}L/RC - \dot{L}/R$$
$$0 < \epsilon_5 \leqq 1 + \dot{R}L/R^2$$

Even in this simple case, it is difficult to get sharp (i.e., almost necessary) sufficient conditions without considering in detail the functions $L(t)$, $C(t)$, and $R(t)$.

7.3 Stability of Nonlinear Systems. In this case, even more than before, intuition must be used to obtain suitable Lyapunov functions. There are a number of well-known examples of how to do this; knowledge of the examples helps to obtain new Lyapunov functions.

Example 9. Linearization About Equilibrium. It is common usage in engineering to consider only small deviations from the operating point (i.e., equilibrium state). This is done by expanding $\mathbf{f}$ in (5-F) in a Taylor series about the operating point and neglecting all higher-order terms.

It is easy to give a rigorous proof of the legitimacy of this procedure by means of the second method. Let $\mathbf{y} = \mathbf{x} - \mathbf{x}_e$ and assume that $\mathbf{f}$ in (5-F) is analytic in a neighborhood of $\mathbf{x}_e$. In that neighborhood $\mathbf{f}$ has a Taylor series which can be written as:

$$d\mathbf{y}/dt = \mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}_e, t) + [\mathbf{F}(t) + \mathbf{G}(\mathbf{y}, t)](\mathbf{x} - \mathbf{x}_e)$$
$$= [\mathbf{F}(t) + \mathbf{G}(\mathbf{y}, t)]\mathbf{y} \tag{38}$$

where $\mathbf{F}(t)$ is the Jacobian matrix of $\mathbf{f}$ evaluated at $\mathbf{x}_e$, t; and $\mathbf{G}(\mathbf{y}, t)$ is a matrix such that $\|\mathbf{G}(\mathbf{y}, t)\mathbf{y}\|/\|\mathbf{y}\|$ tends to 0 with $\|\mathbf{y}\| \to 0$.

Now suppose that

(i) $\|\mathbf{F}(t)\| \leqq c_1$ *for all* t

(ii) *The linearized system*

$$d\mathbf{y}/dt = \mathbf{F}(t)\mathbf{y} \tag{39}$$

is uniformly asymptotically stable.

Then by part (E) of Theorem 3 there is a Lyapunov function $V(\mathbf{y}, t)$ for (39). The derivative of this function along motions of (38) is

$$\dot{V}(\mathbf{y}, t) = -\|\mathbf{y}\|^2_{\mathbf{Q}(t)} + 2\mathbf{y}'\mathbf{P}(t)\mathbf{G}(\mathbf{y}, t)\mathbf{y}$$

By Schwarz's inequality and (16),

$$\leqq -c_0\|\mathbf{y}\|^2 + 2\|\mathbf{P}(t)\mathbf{y}\| \cdot \|\mathbf{G}(\mathbf{y}, t)\mathbf{y}\|$$

By (23), (76), and (80), see Appendix,

$$\leqq -c_0\|\mathbf{y}\|^2 + \frac{c_1^2 c_{10}}{c_0}\|\mathbf{y}\| \cdot \|\mathbf{G}(\mathbf{y}, t)\mathbf{y}\| \tag{40}$$

If we assume further that

(iii) $\|\mathbf{G}(\mathbf{y}, t)\mathbf{y}\|/\|\mathbf{y}\| \to 0$ *uniformly in* t *with* $\|\mathbf{y}\| \to 0$

then (see Theorem 1) $V(\mathbf{y}, t)$ will be a Lyapunov function also for system (38) in a sufficiently small neighborhood of the origin. Therefore we conclude that:

If the linearized system (39) near equilibrium is uniformly asymptotically stable and is a uniformly good approximation to the original system near $\mathbf{x}_e$, *then* $\mathbf{x}_e$ *of (38) is (locally) uniformly asymptotically stable.*

This result includes in it a method of considering small nonlinear terms in an otherwise linear system. This will be discussed in more detail in Example 15.

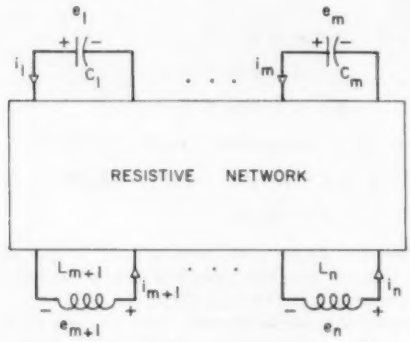


Fig. 11 General RLC network

Example 10. Passive Nonlinear Network. Consider a network of nonlinear inductors, capacitors, and resistors, without internal sources, mutual inductances, or ideal transformers. Let the first m elements be inductors, the next n-m elements capacitors, the rest resistors. See Fig. 11. Let e_j, i_j $(j = 1, \ldots, n)$ be the voltages across and the currents through the energy storage elements with polarities as shown in Fig. 11. Using ψ to denote flux and q to denote charge, the mathematical description of the nonlinear inductors and capacitors is

$$i_j = f_j(\psi_j), \qquad e_j = d\psi_j/dt \qquad (j = 1, \ldots, m)$$

$$e_j = f_j(q_j), \qquad i_j = dq_j/dt \qquad (j = m+1, \ldots, n)$$

Evidently $\psi_1, \ldots, \psi_m, q_{m+1}, \ldots, q_n$ is a natural choice for the state variables $x_1, \ldots, x_n$. The stored energy in the system is given by

$$\begin{aligned} E = V(\mathbf{x}) &= \sum_{j=1}^{n} \int_0^{\infty} e_j i_j dt \\ &= \sum_{j=1}^{m} \int_0^{\psi_j} f_j(\psi_j) d\psi_j + \sum_{j=m+1}^{n} \int_0^{q_j} f_j(q_j) dq_j \\ &= \sum_{j=1}^{n} \int_0^{x_j} f_j(x_j) dx_j \end{aligned} \qquad (41)$$

Assuming the network has an equilibrium state at $\mathbf{x} = \mathbf{0}$, i.e.

$$f_j(0) = 0 \qquad (j = 1, \ldots, n)$$

(41) will be a Lyapunov function if

$$x_j f_j(x_j) > 0, \qquad \int_0^{\infty} f_j(x_j) dx_j = \infty \qquad (j = 1, \ldots, n)$$

and

$$\dot{V}(x) = \sum_{j=1}^{n} e_j i_j < 0$$

The latter condition is always satisfied by passive resistive networks; in other words, by networks containing only resistors whose conductance function is $i = g(e)$ such that (i) $g(0) = 0$, (ii) $g(-e) = -g(e)$, (iii) $g(e)$ is monotonically increasing with e. See equations (6) and (6+) of Reference [30]. Thus *every passive network is asymptotically stable in the large.* While this fact may seem to be obvious physically, the simplest rigorous proof is supplied by the second method of Lyapunov.

An interesting and at present largely unsolved problem which arises in connection with nonlinear control systems is the following: Suppose first that all nonlinearities are functions of a single variable only. Let each nonlinearity $g(\xi)$ be replaced formally by ("linearized") $g(\xi) = K(\xi)\xi$ where either

$$K(\xi) = g(\xi)/\xi \qquad (42)$$

or

$$K(\xi) = dg(\xi)/d\xi = g'(\xi) \qquad (43)$$

where $K(\xi)$ is treated formally as a constant. Assuming stationarity, this leads in general to a differential equation of the form

$$d\mathbf{x}/dt = \mathbf{F}(\mathbf{x})\mathbf{x} \qquad (44)$$

Now suppose the matrix $\mathbf{F}(\mathbf{x})$ is asymptotically stable for all $\mathbf{x}$. Can we conclude from this that the original equation $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is globally asymptotically stable? This has been widely conjectured in engineering circles for some time (in case of linearization (42) by Aizerman [31], in case of linearization (43) by Kalman [32]), and is probably true in most cases. However, no rigorous proof is available at present. More exactly, it is not known in precisely what cases and for precisely what reasons the conjectures fail. There are, however, a number of interesting partial results.

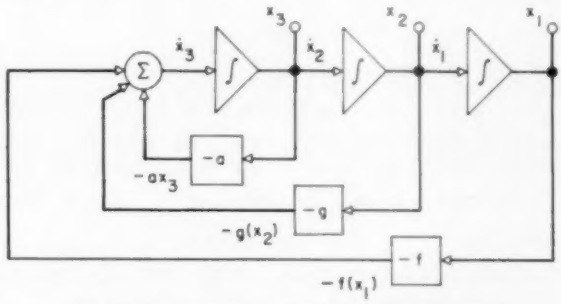


Fig. 12 System corresponding to Barbashin's theorem

Example 11. Stability of a Third-Order Nonlinear System. Barbashin [33, 34] considers the following system (see Fig. 12):

$$\left.\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -f(x_1) - g(x_2) - a_2 x_3 \end{aligned}\right\} \qquad (45)$$

where $f(0) = g(0) = 0$ and f, g are differentiable functions.

Using the "second method" in the form of Corollaries 1.2–1.3, Barbashin shows:

The equilibrium state $\mathbf{x}_e = \mathbf{0}$ *of system (45) is asymptotically stable in the large if:*

(i) $a_2 > 0$
(ii) $f(x_1)/x_1 \geqq \epsilon_1 > 0, \qquad x_1 \neq 0$
(iii) $a_2 g(x_2)/x_2 - f'(x_1) \geqq \epsilon_2 > 0, \qquad x_2 \neq 0$

If we replace $f(x_1)/x_1$ and $f'(x_1)$ by a_0 and $g(x_2)/x_2$ by a_1, then the foregoing conditions become:

(i′) $a_2 > 0$ (ii′) $a_0 > 0$ (iii′) $a_2a_1 - a_0 > 0$

which are just the Routh-Hurwitz conditions for the linear system

$$\dddot{x}_1 + a_2\ddot{x}_1 + a_1\dot{x}_1 + a_0x_1 = 0$$

Notice, however, that conditions (i–iii) above involve *both* kinds of linearizations mentioned in the foregoing.

Let

$$F(x_1) = \int_0^{x_1} f(x_1)dx_1 \quad \text{and} \quad G(x_2) = \int_0^{x_2} g(x_1)dx_2$$

and define as a candidate for a Lyapunov function of (45):

$$V(\mathbf{x}) = a_2F(x_1) + f(x_1)x_2 + G(x_2) + (a_2x_2 + x_3)^2/2 \quad (46)$$

In view of (ii), $f'(0) > 0$, and $F(x_1) > 0$ when $x_1 \neq 0$. Further, by (i) and (iii), it follows that $G(x_2) > 0$ when $x_2 \neq 0$. When $x_2 = 0$ (46) is obviously nonnegative and 0 only if $x_1 = x_3 = 0$. When $x_2 \neq 0$, the first three terms in (46) can be written as

$$\frac{[2G(x_2) + f(x_1)x_2]^2 + 4\int_0^{x_1} f(x_1)\int_0^{x_2} [a_2g(x_2)/x_2 - f'(x_1)]x_2dx_2dx_1}{4G(x_2)}$$

The first term in the foregoing is always positive. By (iii) the inner integral above is positive, and therefore by (ii) the outer integral is nonnegative. Putting such facts together shows that V is positive unless $\mathbf{x} = \mathbf{0}$. Because all integrands in question are uniformly bounded from below by (ii) and (iii), it follows further that $V(\mathbf{x}) \to \infty$ when $\|\mathbf{x}\| \to \infty$. Hence V satisfies conditions (i-A), (iii-A) and (iv-A) of Corollary 1.2. After a short calculation, it is found that

$$\dot{V}(\mathbf{x}) = -[a_2g(x_2)/x_2 - f'(x_1)]x_2^2$$

By (iii), $\dot{V}(\mathbf{x}) = 0$ only if $x_2 = 0$. Using exactly the same argument as in Example 7, it follows that $\dot{V}(\phi(t;\ \mathbf{x}_0,\ 0))$ does not vanish identically along any trajectory starting at $\mathbf{x}_0 \neq \mathbf{0}$. This verifies hypotheses (ii_1-A), (ii_2-A) of Corollary 1.3 which completes the proof that (46) is a Lyapunov function for (45).

Perhaps the best general result on stability of nonlinear systems (not restricted by dimensionality as in the foregoing example) is the following:

Theorem 4. (Krasovskii [35]) *Consider the continuous-time, free, stationary dynamic system:*

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}) \qquad (\mathbf{f}(\mathbf{0}) = \mathbf{0}) \qquad (5\text{-FS})$$

ASSUME *that* $\mathbf{f}$ *has continuous first partial derivatives and that its Jacobian matrix* $\mathbf{F}(\mathbf{x}) = [\partial f_i/\partial x_j]$ *satisfies the condition* ($\epsilon > 0$ *is an arbitrary constant*):

(i) $\hat{\mathbf{F}}(\mathbf{x}) = [\mathbf{F}(\mathbf{x}) + \mathbf{F}'(\mathbf{x})] \leqq -\epsilon\mathbf{I} < \mathbf{0}$

THEN *the equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *of the system* (5-FS) *is asymptotically stable in the large and*

$$V(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2$$

is one of its Lyapunov functions.

PROOF. Computing $\dot{V}$, we find that

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}'(\mathbf{grad}\ V)\\ &= 2\sum_{j=1}^{n}\dot{x}_j\sum_{i=1}^{n}f_i(x)\frac{\partial f_i(x)}{\partial x_j}\\ &= 2\mathbf{f}'(\mathbf{x})\mathbf{F}(\mathbf{x})\dot{\mathbf{x}}\\ &= \|\mathbf{f}(\mathbf{x})\|^2\hat{\mathbf{F}}_{(\mathbf{x})} \leqq -\epsilon\|\mathbf{x}\|^2 < 0\end{aligned}$$

the last step following by hypothesis (i). Thus $\dot{V}$ is negative definite.

It remains to show that V is positive definite and tends to ∞ with $\|\mathbf{x}\| \to \infty$.

Let $\mathbf{c}$ be any constant, nonzero vector. The set of vectors $\{\alpha\mathbf{c};\ 0 \leqq \alpha \leqq 1\}$ is a straight line (ray) connecting the origin with $\mathbf{c}$. Integrating along this ray, we have the identity

$$f_i(\mathbf{c}) = \sum_{j=1}^{n}\left[\int_0^1 c_j\frac{\partial f_i(\alpha\mathbf{c})}{\partial x_j}d\alpha\right] \qquad (46)$$

Suppose that $\mathbf{f}(\mathbf{c})$ is $\mathbf{0}$ for some $\mathbf{c} \neq \mathbf{0}$. Then obviously

$$\begin{aligned}0 = \mathbf{c}'\mathbf{f}(\mathbf{c}) &= \sum_{i=1}^{n}c_if_i(\mathbf{c})\\ &= \int_0^1\left[\sum_{i,j=1}^{n}c_i\frac{\partial f_i(\alpha\mathbf{c})}{\partial x_j}c_j\right]d\alpha\\ &\leqq -\epsilon\|\mathbf{c}\|^2/2 < 0\end{aligned}$$

which is a contradiction. Hence V is positive definite. The preceding argument shows also that $\mathbf{x}'\mathbf{f}(\alpha\mathbf{x}) \to -\infty$ with $\alpha \to \infty$, for any fixed vector $\mathbf{x} \neq \mathbf{0}$. But this can happen only if at least one component of $\mathbf{f}(\mathbf{x})$ tends to infinity in absolute value as $\|\mathbf{x}\|$ tends to infinity which completes the proof.

The most obvious illustration of this result is the following:

Example 12. Passive Linear Stationary System. In Example 10, in the linear case, one can define the state variables in such a way that the infinitesimal transition matrix $\mathbf{F}$ is negative definite. This is due to the fact that the terminal impedance matrix of an n-port network containing only linear, passive resistors is negative definite. Then stability is immediate, by Theorem 4. In this special case, therefore, the Routh-Hurwitz inequalities are actually identical with the ones obtained by testing for negative definiteness.

A less trivial example is:

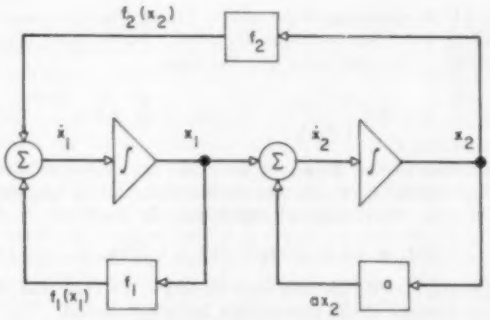


Fig. 13 System illustrating Krasovskii's theorem

Example 13. Second-Order System With Two Nonlinearities Fig. 13 illustrates the system

$$\begin{aligned}\dot{x}_1 &= f_1(x_1) + f_2(x_2)\\ \dot{x}_2 &= x_1 + ax_2\end{aligned} \qquad (47)$$

where $f_1(0) = f_2(0) = 0$; f_1 and f_2 are differentiable functions. Here

$$\hat{\mathbf{F}}(\mathbf{x}) = \begin{bmatrix}2f_1'(x_1) & 1 + f_2'(x_2)\\ 1 + f_2'(x_2) & 2a\end{bmatrix}$$

To satisfy the hypothesis of Theorem 4 we must have

$$4af_1'(x_1) - [1 + f_2'(x_2)]^2 \geq \epsilon^2 > 0 \quad \text{for all } \mathbf{x}$$
$$f_1'(x_1) \leq -\epsilon < 0 \text{ for all } x_1 \tag{48}$$

These requirements are considerably more restrictive than the formal Routh-Hurwitz conditions which would be

$$af_1(x_1)/x_1 - f_2(x_2)/x_2 \geq \epsilon^2 > 0 \quad \text{for all } x_1 \neq 0, \quad x_2 \neq 0$$
$$f_1(x_1)/x_1 + a \leq -\epsilon < 0 \quad \text{for all } x_1 \neq 0 \tag{49}$$

or, using the other type of linearization,

$$af_1'(x_1) - f_2'(x_2) \geq \epsilon^2 > 0 \quad \text{for all } \mathbf{x}$$
$$f_1'(x_1) + a \leq -\epsilon < 0 \quad \text{for all } x_1 \tag{50}$$

where (50) evidently implies (49). Using a special Lyapunov function, Krasovskii [36] was able to show that the asymptotic stability in the large of the origin of the system (47) is guaranteed even when conditions (48) are replaced by the *first* of conditions (50) and the *second* of conditions (49). On the other hand, as Krasovskii also shows, conditions (49) are not sufficient even for local asymptotic stability.

It is not a simple matter to obtain Routh-Hurwitz-like conditions for nonlinear systems since it is not clear what *kind* of linearization is the most natural one. Much more mathematical research remains to be done on these problems.

8—Estimation of Transient Behavior

If the stability of a given system has been established by means of a Lyapunov function, the latter can be used to estimate the rapidity of transient response, effect of perturbations or variation in parameters, etc. All of these applications stem from regarding the Lyapunov function as a measure of distance in the state space.

Consider the obvious inequality

$$\dot{V}(\mathbf{x}, t) = [\dot{V}(\mathbf{x}, t)/V(\mathbf{x}, t)]V(\mathbf{x}, t)$$
$$\leq -\eta V(\mathbf{x}, t) \tag{51}$$

where η is the minimum of the ratio $-\dot{V}(\mathbf{x}, t)/V(\mathbf{x}, t)$ in some region of the state space excluding the origin. For instance, in case of uniform asymptotic stability in the large,

$$\eta = \operatorname*{Inf}_{\mathbf{x}} \left\{ \frac{\gamma(\|\mathbf{x}\|)}{\beta(\|\mathbf{x}\|)};\ 0 < \|\mathbf{x}\| \leq r \right\} \geq 0\ ; \tag{52}$$

if r is chosen so that $\beta(\|\mathbf{x}_0\|) < \alpha(r)$ then the motion will never leave the region $\|\mathbf{x}\| \leq r$ so that the inequality (51) is applicable for all $t \geq t_0$. By elementary calculation, (51) leads to

$$V(\phi(t;\ \mathbf{x}_0, t_0), t) \leq \exp\ [-\eta(t - t_0)]V(\mathbf{x}_0, t_0) \tag{53}$$

Interpreting V as the distance from the origin, it is clear that (53) gives an estimate of how fast equilibrium is approached.

In fact, η^{-1} can be interpreted as the largest time-constant over a certain region in state space and may be regarded as a figure of merit of the system. A number of points should be observed in this connection:

(*a*) If $\dot{V}$ is only semidefinite, we may be able to conclude asymptotic stability by means of Corollary 1.3. But then $\eta = 0$ and we cannot use (51) to estimate the rate of approach to equilibrium—in fact, then (51) merely implies stability. Fortunately, by Theorem 2, one can always find, in principle, a Lyapunov function with negative definite $\dot{V}$ in case of uniform asymptotic stability. Another situation where semidefinite $\dot{V}$ leads to difficulty occurs in the next section.

(*b*) It may be useful to consider the values of η over several different regions of state space. For instance, in Example 2, $\eta = 2aV$ by equation (4). Since $V \to 0$ with $\mathbf{x} \to \mathbf{0}$, it follows that, if η is defined by (52), it is zero. On the other hand, in any annular region $r_1 \leq \|\mathbf{x}\| \leq r_2$, $\eta = 2ar_1^2$.

(*c*) The specific value of η depends on the Lyapunov function chosen. Therefore one may want to choose V so as to maximize η. Little is known at present as to how this can be done.

Example 14. Calculation of η in Linear Stationary Case. Consider a free, linear, stationary, asymptotically stable system. Let $\mathbf{Q}$ be an arbitrary symmetric, positive-definite matrix and find the corresponding symmetric, positive-definite matrix $\mathbf{P}$ by Corollary 3.1. Then

$$V(\mathbf{x}) = \|\mathbf{x}\|_{\mathbf{P}}^2, \qquad \dot{V}(\mathbf{x}) = -\|\mathbf{x}\|_{\mathbf{Q}}^2 \tag{54}$$

A definition of η analogous to (52) is then

$$\eta = \operatorname*{Min}_{\mathbf{x}} \{\|\mathbf{x}\|_{\mathbf{Q}}^2;\ \|\mathbf{x}\|_{\mathbf{P}}^2 = 1\} \tag{55}$$

and now the minimization can be performed with the help of the usual Lagrange multiplier technique. Let Γ be the Lagrange multiplier; then minimizing $\mathbf{x}'(\mathbf{Q} - \Gamma\mathbf{P})\mathbf{x}$ with respect to $\mathbf{x}$ implies that the minimum occurs at a value $\mathbf{x}^*$ of $\mathbf{x}$ such that

$$(\mathbf{Q} - \Gamma\mathbf{P})\mathbf{x}^* = \mathbf{0} \tag{56}$$

Therefore

$$\mathbf{x}^{*\prime}\mathbf{Q}\mathbf{x}^* = \Gamma\mathbf{x}^{*\prime}\mathbf{P}\mathbf{x}^* = \Gamma > 0$$

which is a minimum if Γ is. But, by (56), Γ is an eigenvalue of the matrix $\mathbf{QP}^{-1}$. Hence

$$\eta = \lambda_{\min}(\mathbf{QP}^{-1}) \tag{57}$$

The inputs or perturbations acting on a control system are frequently not known accurately, although some estimate of their maximum amplitude may be available. Similarly, the function $\mathbf{f}$ may not be known exactly, or may change due to parameter variations. Both effects may be represented by adding a term on the right-hand side of (5-F). In such cases, there may be no asymptotic stability but one can ask the question, "What is the smallest spherical neighborhood about the origin which the motions of the perturbed system are *sure* to enter?" The radius of this neighborhood depends on the bounds of the perturbation and the parameter variation and can be easily estimated using the "second method."

Write (5-F) in the form

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, t) + \mathbf{g}(\mathbf{x}, t),$$

where

$$\|\mathbf{g}(\mathbf{x}, t)\| \leq c_0 < \infty \tag{58}$$

Let $V(\mathbf{x}, t)$ be a Lyapunov function for the system (5-F) in accordance with Theorem 1, with derivative $\dot{V}(\mathbf{x}, t)$. Then the derivative of V for the system (58) is

$$\dot{V}(\mathbf{x}, t) + \mathbf{g}'(\mathbf{x}, t)[\mathbf{grad}\ V(\mathbf{x}, t)]$$
$$\leq -\gamma(\|\mathbf{x}\|) + c_0\|\mathbf{grad}\ V(\mathbf{x}, t)\| \tag{59}$$

If V and γ are given in manageable form, one can then estimate a constant $c_1(c_0)$ such that if $\|\mathbf{x}\| > c_1$, then expression (59) is negative.

Example 15. Effect of Perturbations and Parameter Variations in the Linear Stationary Case. Consider the system

$$d\mathbf{x}/dt = \mathbf{F}\mathbf{x} + \mathbf{G}(\mathbf{x})\mathbf{x} + \mathbf{u}(t) \tag{60}$$

where

$$\|\mathbf{G}(\mathbf{x})\| \leq c_0$$
$$\|\mathbf{u}(t)\| \leq c_1$$

Assume that $\mathbf{F}$ is asymptotically stable and define a Lyapunov

function by means of Corollary 3.1, as in Example 14, equation (54). Then

$$\dot{V}(\mathbf{x}, t) = -\|\mathbf{x}\|^2_{\mathbf{Q}} + 2[\mathbf{x}'\mathbf{G}'(\mathbf{x}) + \mathbf{u}'(t)]\mathbf{P}\mathbf{x} \tag{61}$$

By Schwarz's inequality,

$$|\mathbf{x}'\mathbf{G}'(\mathbf{x})\mathbf{P}\mathbf{x}| \leq c_0\|\mathbf{x}\| \cdot \|\mathbf{P}\mathbf{x}\|$$

and

$$|\mathbf{u}'(t)\mathbf{P}\mathbf{x}| \leq c_1\|\mathbf{P}\mathbf{x}\|$$

Hence (61) becomes

$$\dot{V}(\mathbf{x}, t) \leq -\|\mathbf{x}\|^2_{\mathbf{Q}} + 2c_0\|\mathbf{x}\| \cdot \|\mathbf{P}\mathbf{x}\| + 2c_1\|\mathbf{P}\mathbf{x}\| \tag{62}$$

Now observe that (proof as in Example 14)

$$\|\mathbf{P}\mathbf{x}\|^2 = \|\mathbf{x}\|^2_{\mathbf{P}^2} \leq \lambda_{\max}(\mathbf{P})\|\mathbf{x}\|^2_{\mathbf{P}}$$

Using this, it follows,

$$\dot{V}(\mathbf{x}, t) \leq [-\eta + 2c_2(c_0)]V + 2c_2(c_1)V^{1/2} \tag{63}$$

where η is as defined in Example 14,

$$c_2(c_0) = c_0\lambda_{\max}{}^{1/2}(\mathbf{P})/\lambda_{\min}{}^{1/2}(\mathbf{P})$$

and

$$c_2(c_1) = c_1\lambda_{\max}{}^{1/2}(\mathbf{P})$$

Therefore if

$$\|\mathbf{x}\| \geq \|\mathbf{x}\|_{\mathbf{P}}/\lambda_{\max}{}^{1/2}(\mathbf{P})$$

$$\geq 2c_1[\eta - 2c_2(c_0)]^{-1} \tag{64}$$

then $\dot{V}$ is negative.

The quantity $[\eta - 2c_2(c_0)]^{-1}$ can be interpreted as the upper bound on the time constant taking into account the parameter variations. The upper bound of the radius of the region which the motions are sure to enter is proportional to this quantity and to the bound on the noise $\mathbf{u}(t)$.

9—Relations With System Optimization

An important aspect of the "second method" of Lyapunov is its close relationship with problems in the dynamic optimization of control systems. This interrelation has not been explored so far in detail even in the Russian literature and provides much incentive for future research.

Briefly, the situation is as follows. Let $\rho(\mathbf{x})$ be a continuous, nonnegative function which serves as the error criterion of a regulator whose purpose is to maintain the system at all times as close as possible to the equilibrium state $\mathbf{x}_e = \mathbf{0}$. Thus $\rho(\mathbf{0}) = 0$. For instance, one could take $\rho(\mathbf{x}) = x_1^2$, or $\rho(\mathbf{x}) = 1 - \exp(-x_1^2)$, or $\rho(\mathbf{x}) = \sum_{i=1}^{n} |x_i|$, etc.

Now define the *performance index* of the system as the integrated error criterion (assuming stationarity):

$$V(\mathbf{x}) = \int_0^\infty \rho(\phi(\tau;\ \mathbf{x}, 0))d\tau \tag{65}$$

This integral is a positive number and can serve as a figure of merit of the system. (When the integral is infinite, it is useless; this can be easily avoided by replacing the upper limit of integration by T, where T denotes an interval of time over which the behavior of the system is of interest.)

If the regulator contains adjustable parameters $\mathbf{p}$ these can be picked in such a way that the value of $V(\mathbf{x}, \mathbf{p})$ is minimized with respect to those parameters for one or more initial state $\mathbf{x}$. This formulation of the optimization problem contains the well-known Wiener-Hall theory [37].

Now the mere fact that the integral (65) is minimized by a physically realizable system does not imply at all that this system is also asymptotically stable. The latter will be the case, however, if $V(\mathbf{x})$ is a suitable Lyapunov function.

To be more precise, consider only linear stationary systems (5-LS). Because ρ is nonnegative and the integrand is continuous in τ, the integral in (65) can be zero only if

$$\rho(\phi(\tau;\ \mathbf{x}, 0)) = 0 \quad \text{for all } \tau \geq 0 \tag{66}$$

Assume now that $\rho(\mathbf{x})$ is positive definite. Then (66) implies, using also linearity and stationarity:

$$\phi(\tau;\ \mathbf{x}, 0) = (\exp \mathbf{F}\tau)\mathbf{x} = \mathbf{0},\ \tau \geq 0$$

Since the transition matrix is nonsingular, this implies $\mathbf{x} = \mathbf{0}$, which shows that V is positive definite.

Assuming further that V is *finite* in some neighborhood $\|\mathbf{x}\| \leq r$ of the origin, a simple calculation (as in (24)) shows that in the same neighborhood

$$\dot{V}(\mathbf{x}) = -\rho(\mathbf{x}) < 0 \quad \text{when } \mathbf{x} \neq \mathbf{0}$$

i.e., $\dot{V}$ is negative definite. Therefore under the assumptions V satisfies requirements (i-A), (ii-A) of Corollary 1.2 so that (5-LS) is asymptotically stable; by linearity, it is then also asymptotically stable in the large so that the restriction about a finite neighborhood of the origin made above is immaterial. Thus we have obtained the following simple but useful result [16]:

Theorem 5. (Kalman) *Consider a free, linear, stationary dynamic system with an equilibrium state at the origin and* ASSUME

(i) *The error criterion $\rho(\mathbf{x})$ is positive definite and $\rho(\mathbf{0}) = 0$;*
(ii) *The performance index $V(\mathbf{x})$ defined by (65) is finite in some neighborhood of the origin.*

THEN $\mathbf{x}_e = \mathbf{0}$ *is asymptotically stable.*

The assumptions on $\rho(\mathbf{x})$ could be considerably weakened, of course; the extension of the theorem to the linear nonstationary case is completely straightforward.

The importance of this theory is emphasized by the following facts:

(*a*) In the conventional theory of optimization the assumption of positive definiteness of ρ is missing. For this reason, the theory in its classical form is unsatisfactory since there is no proof in general that the optimized system is stable. In the simple cases usually discussed in textbooks, this omission is actually not serious due to the *accidental* fact (cf. here Examples 7 and 11) that often even positive semidefinite error criteria imply that $V(\mathbf{x})$ is a Lyapunov function since the error criterion does not vanish identically along any trajectory (Corollary 1.3). But this is much less likely to happen in systems with several inputs and outputs—one of the reasons why relatively little progress has been made in the theory of multiloop systems. These matters are discussed in detail in a recent paper of Kalman [16, 41].

(*b*) If $V(\mathbf{x})$ can be made finite by *some* choice of the parameters of control system, then it follows that the optimal system will also be stable since $V_{\text{opt}}(\mathbf{x}) \leq V(\mathbf{x})$. But if the open-loop system is unstable, it is by no means obvious that $V_{\text{opt}}(\mathbf{x})$ can be made finite for all $\mathbf{x}$ and thereby stabilize the system. In fact, in some nonlinear systems [38–40], one can at best get local stability in this case. However, Kalman has shown [16, 41] that, if the definition of the control problem is physically meaningful at all, then in the *linear* case an originally unstable system can always be stabilized by a suitably designed controller.

(*c*) Note that Theorem 5 is not based on any special assumptions concerning the error criterion ρ, for instance, that it should be a quadratic function of the state. From the point of view of getting general results, the particular form of ρ is unimportant; unfortunately, however, the problem of minimizing $V(\mathbf{x}, \mathbf{p})$ is

usually difficult to carry out and leads to a nonlinear feedback system whenever ρ is not quadratic.

(*d*) By minimizing $V(\mathbf{x}, \mathbf{p})/\rho(\mathbf{x})$ with respect to $\mathbf{p}$, one automatically obtains the best possible upper bound on the time constant of the system as discussed in Section 8.

(*e*) When $\rho(\mathbf{x}) = x_1^2$, the classical method of evaluation of (65) in the linear stationary case is based on the identity

$$\int_0^\infty x_1^2(\tau)d\tau = \frac{1}{2\pi}\int_{-i\infty}^{i\infty} X_1(i\omega)X_1(-i\omega)d\omega$$

where $X_1(i\omega)$ is the Fourier transform of $x_1(t)$. The right-hand side of the integral is then evaluated by means of the calculus of residues; the resulting formulas have been tabulated [37]. Some may regard this as a clever application of the calculus of residues. However, the same task can be solved in a much more general and elementary way by applying Corollary 3.1, i.e., by formulating the problem in the "right" way from the stated point of view and then solving the linear algebraic equations (19). Thus Corollary 3.1 can be used not only as a method of checking stability but also for system optimization. The following elementary example, suggested by Herschel [42], illustrates this:

Example 16. Conventional Optimization of Second-Order Stationary Systems. Consider the second-order linear stationary system in the output x_1:

$$\ddot{x}_1 + 2\zeta\dot{x}_1 + x_1 = 0$$

or

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= -x_1 - 2\zeta x_2\end{aligned} \tag{67}$$

where ζ is an arbitrary constant.

The problem is to minimize the performance index

$$V(\mathbf{x}, \zeta) = \|\mathbf{x}\|^2_{\mathbf{P}} = \int_{t_0}^\infty \|\phi(\tau;\ \mathbf{x},\ t_0)\|^2_{\mathbf{Q}}d\tau$$

where

$$\mathbf{Q} = \begin{bmatrix}1 & 0\\ 0 & \mu\end{bmatrix},\quad \mu > 0$$

Using Corollary 3.1, the solution of equation (19) is

$$\mathbf{P}(\zeta) = \begin{bmatrix}\zeta + \dfrac{1+\mu}{4\zeta} & 1/2\\ 1/2 & \dfrac{1+\mu}{4\zeta}\end{bmatrix}$$

or

$$V(\mathbf{x}, \zeta) = \zeta x_1^2 + \left[\frac{1+\mu}{4\zeta}\right](x_1^2 + x_2^2) + x_1x_2 \tag{68}$$

Suppose now that the constant $\zeta(\mu)$ is chosen in such a way that V is a minimum when $x_2 = 0$. This leads to

$$\zeta^*(\mu) = (\sqrt{1+\mu})/2$$

as the optimal choice of ζ.

Notice that $\zeta^*(0) = 0.5$, a well-known result. The objection has been frequently raised that this type of optimization, (i.e., using $\rho(\mathbf{x}) = x_1^2$) yields a system which is too oscillatory. The usual "rule of thumb" is given to be $\zeta \cong 0.7$. This value is given by $\zeta^*(1) = 1/\sqrt{2} = 0.707$. Thus the use of a *positive-definite* error criterion answers some of the objections raised against optimization based on the integrated squared error.

If ζ could be varied instantaneously, then we could just as well write system (67) in the form

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= -x_1 + u_2(t)\end{aligned}$$

where $u_2(t) = \zeta(t)x_2$. If there are no limits on $\zeta(t)$ then there are also no limits on $u_2(t)$ and then Kalman's theory of optimization [16] (see also reference [15]) applies. It follows that $V(\mathbf{x})$ for this system can be made arbitrarily small by letting $\mathbf{u}(t)$ be some linear function of the state:

$$u_2^*(t) = a_1x_1(t) + a_2x_2(t)$$

and therefore we would have also

$$\zeta^*(t) = a_1x_1(t)/x_2(t) + a_2$$

But even if ζ must be a constant always, it follows that the foregoing optimization procedure is not satisfactory, because one should conceivably still select ζ *as a function of the initial state* $\mathbf{x}$. For instance, if the initial state is $x_1 = 0$ and x_2 arbitrarily, then the optimum value of ζ for this initial state is ∞.

Example 17. Strict Optimization of Second-Order Stationary Systems. A more satisfactory approach to the problem of the preceding example is to regard the optimization as a (two-person, zero-sum) *game* between Nature and Man. Nature chooses the initial state $\mathbf{x}$; Man chooses a (constant) value of ζ. Nature attempts to maximize and Man to minimize the time constant:

$$\tau = \eta^{-1} = V(\mathbf{x}, \zeta)/\rho(\mathbf{x})$$

It is intuitively obvious and easily proved [43] that if Nature plays first, the *payoff* (i.e., the value of τ) will be at least as great as it would be if Man were to play first. This is expressed by the well-known inequality [43]:

$$\tau_* = \max_{\mathbf{x}}\{\min_{\zeta}[V(\mathbf{x}, \zeta)/\rho(\mathbf{x})]\} \leqq \leqq \min_{\zeta}\{\max_{\mathbf{x}}[V(\mathbf{x}, \zeta)/\rho(\mathbf{x})]\} = \tau^* \tag{69}$$

Here τ_* is the minimum that Nature is sure to "get" and τ^* is the maximum that Man may have to "pay."

The payoff function (68–69) of this continuous game is *convex* in the *strategies* ζ; for this class of games a fairly complete theory is available (Reference [43], Chap. 12). It can be easily computed that here actually $\tau^* = \tau^*$. This means that the game has a *saddle point*, i.e., a *solution* of the game consists of *pure strategies* $\mathbf{x}^*$ and ζ^* which are the values of $\mathbf{x}$, ζ at which the Min-Max (or Max-Min) of $V(\mathbf{x}, \zeta)/\rho(\mathbf{x})$ is attained. An elementary but somewhat lengthy calculation shows that, when $\mu = 1$,

$$\begin{aligned}x_1^* &= \sqrt{(1+\sqrt{5})}\|\mathbf{x}\|/2 = 0.899\|\mathbf{x}\|\\ x_2^* &= \sqrt{(3-\sqrt{5})}\|\mathbf{x}\|/2 = 0.437\|\mathbf{x}\|\\ \zeta^* &= \sqrt{\frac{2}{1+\sqrt{5}}} = 0.786\\ \tau_* &= \tau^* = 1/2(\zeta^* + 1/\zeta^* + \sqrt{\zeta^{*2} + 1}) = 1.665\end{aligned}$$

Since the value of the game does not change if $\mathbf{x}$ is replaced by $\alpha\mathbf{x}$ ($\alpha \neq 0$), the optimal strategy $\mathbf{x}^*$ is determined only up to a constant multiplier.

Thus we have a solution of the optimization problem which does not depend on any arbitrary assumptions as to the initial state; τ^* is the minimum of the maximum possible values of the time constant. This agrees well with the "rule of thumb" quoted above.

Such problems have not been considered in the literature so far and it seems to be an open question as to whether the optimal

solution of games of this type always results in pure strategies.

10—Applications to Design

The "second method" can be used in a variety of ways in connection with system design. A particular advantage of this way of approaching design is that one has then a direct way of assuring stability. A full discussion of these matters is given in the books of Lur'e [7] and Letov [8] cited before.

To give an idea of the possibilities present, we discuss briefly the problem of linear compensation of control systems subject to saturation.

Consider a plant governed by the equations

$$d\mathbf{x}/dt = \mathbf{F}\mathbf{x} + \mathbf{D}\mathbf{u}(t) \tag{70}$$

where the control variables are subject to the constraints,

$$|u_i(t)| < a_i < \infty \qquad (i = 1, \ldots, m) \tag{71}$$

This is essentially the relay or saturating servo problem. The problem is to return every initial state to the origin.

It is well known [38–40] that if $\mathbf{F}$ has eigenvalues with positive real parts then there are some states which cannot be returned to the origin by *any* control subject to the constraints (71). Hence it is assumed that the free system in (70) is asymptotically stable; i.e., that $Re\ \lambda_i(\mathbf{F}) < 0$ for all i. The case where $\mathbf{F}$ has eigenvalues with zero real parts can be treated similarly, albeit much less simply.

Under these assumptions, consider the problem of optimizing the transient behavior of (70) by suitable choice of $\mathbf{u}(t)$. A rigorous optimization, such as minimization of an integral performance index (65) or taking any initial state to the origin in minimum time, will lead to $\mathbf{u}(t)$ being a *nonlinear* function of the instantaneous state. This problem is not yet completely solved.

The following method, suggested by Bass [44], leads to a very simple design procedure of practical significance.

Choose $\mathbf{Q} > \mathbf{0}$ arbitrarily, and employ Corollary 3.1 to get $\mathbf{P} > \mathbf{0}$. Then $\|\mathbf{x}\|^2_{\mathbf{P}} = V(\mathbf{x})$ is a Lyapunov function for system (70), with $\mathbf{u}(t)$ identically zero. Now choose $\mathbf{u}(t)$ so as to make $\dot{V}$ as negative as possible. From (70)

$$\dot{V}(\mathbf{x}(t), t) = -\|\mathbf{x}(t)\|^2_{\mathbf{Q}} + 2\mathbf{u}'(t)\mathbf{D}'\mathbf{P}\mathbf{x}(t)$$

The second term in this expression will be a minimum whenever each control variable $u_i(t)$ has its maximum magnitude, with a sign opposite to that of the corresponding component of the vector $\mathbf{D}'\mathbf{P}\mathbf{x}(t)$. Hence

$$u_i^*(t) = -a_i \operatorname{sgn} (\mathbf{D}'\mathbf{P}\mathbf{x}(t))_i \qquad (i = 1, \ldots, m) \tag{72}$$

where *sgn* is as defined in (21).

This method of design is not a true optimum since

$$\operatorname*{Min}_{\mathbf{u}(t)} \{V(\mathbf{x}, \mathbf{u}(t))/[-\dot{V}(\mathbf{x}, \mathbf{u}(t))]\} < V(\mathbf{x}, \mathbf{u}(t)) \cdot \{\operatorname*{Min}_{\mathbf{u}(t)} [1/-\dot{V}(\mathbf{x}, \mathbf{u}(t))]\}$$

On the other hand, the design has the advantage that it leads to a system in which there is only linear feedback preceding the nonlinearity.

The result (72) is deceptively simple and it is important to bear in mind two major mathematical difficulties.

First, if the optimal control defined by (72) is substituted into (70) there is no guarantee that solutions of the differential equation will *exist*, since *sgn* is a discontinuous function. This is a well-known phenomenon in the theory of relay servos which was first discovered by Flügge-Lotz [45] and later studied in considerable detail by André and Seibert [46]. However, the "second method" still applies, after a small modification in the definition of $\dot{V}$; the conclusions so obtained agree with a rigorous treatment of the relay servo problem. The difficulties just mentioned may be avoided by replacing (72) with

$$u_i^* = -a_i \operatorname{sat} k(\mathbf{D}'\mathbf{P}\mathbf{x})_i = u_i^*(k\mathbf{D}'\mathbf{P}\mathbf{x}) \quad (i = 1, \ldots, m) \tag{72-A}$$

where

$$\operatorname{sat} x = \begin{cases} +1, & x > 1 \\ x, & |x| \leq 1 \\ -1, & x < -1 \end{cases}$$

and letting the positive constant k be as large as desired. This provides an arbitrarily good approximation to *sgn* x.

Second, the design represented by (72-A) is realizable only if all state variables can be *measured*. If this is not the case, the unmeasurable state variables must be "synthesized" from the measurable ones by simulating [12] a part of the dynamic equation (70). The necessary simulation equipment (in essence, a small computer) then becomes part of the controller itself (see Example 18 and Fig. 14). However, it is then not at all obvious as to whether the over-all system, including the simulator, will be asymptotically stable. Moreover, there may be difficulties due to inaccurate knowledge of $\mathbf{F}$ and $\mathbf{D}$.

To examine these questions more fully, let $\mathbf{y}$ denote the measurable and $\mathbf{z}$ the unmeasurable state variables. Then (70) can be written in the form:

$$\begin{bmatrix} d\mathbf{y}/dt \\ d\mathbf{z}/dt \end{bmatrix} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} + \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix} \mathbf{u}(t)$$

where $\mathbf{F}_{ij}$ and $\mathbf{D}_i$ $(i, j = 1, 2)$ denote submatrices of $\mathbf{F}$ and $\mathbf{D}$ defined in the obvious way. Further, let $\hat{\mathbf{z}}$ denote the "synthetic" state variables, and $\hat{\mathbf{F}}_{ij}$, $\hat{\mathbf{D}}_i$ the best available estimate of the quantities $\mathbf{F}_{ij}$, $\mathbf{D}_i$. Use the tilde to denote the error in these estimates such as $\tilde{\mathbf{z}} = \mathbf{z} - \hat{\mathbf{z}}$, etc. Then the equations of the" simulator" are:

$$d\hat{\mathbf{z}}/dt = \hat{\mathbf{F}}_{21}\mathbf{y} + \hat{\mathbf{F}}_{22}\hat{\mathbf{z}} + \hat{\mathbf{D}}_2\mathbf{u}(t)$$

which contains only known quantities. (Assuming there are no random disturbances, $\mathbf{u}(t)$ is always measurable.)

The form of the optimal control is still represented by (72-A) but now $\mathbf{D}$ is replaced by $\hat{\mathbf{D}}$ and $\mathbf{P}$ by $\hat{\mathbf{P}}$, the latter being calculated via (19) from $\mathbf{Q}$ and $\mathbf{F}$. Then (with some obvious notation conventions) the over-all system equations become:

$$\left.\begin{aligned} &(a)\ d\mathbf{x}/dt = \hat{\mathbf{F}}\mathbf{x} + \hat{\mathbf{D}}\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}(\mathbf{x} + \tilde{\mathbf{z}})) \\ &\qquad\qquad + \tilde{\mathbf{F}}\mathbf{x} + \tilde{\mathbf{D}}\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}(\mathbf{x} + \tilde{\mathbf{z}})) \\ &(b)\ d\tilde{\mathbf{z}}/dt = \hat{\mathbf{F}}_{22}\tilde{\mathbf{z}} + \tilde{\mathbf{F}}_2 \cdot \mathbf{x} \end{aligned}\right\} \tag{73}$$

Equations (73) define a dynamic system in the state space consisting of all ordered pairs of vectors $(\mathbf{x}, \mathbf{z})$. The two equations are loosely coupled due to errors in the prediction of $\mathbf{z}$ and in the knowledge of $\mathbf{F}$ and $\mathbf{D}$. We assert:

Theorem 6. (Kalman) *The free, stationary dynamic system (73) is asymptotically stable in the large, for arbitrary $k \geq 0$, if the following conditions hold:*

(i) $Re\ \lambda_i(\hat{\mathbf{F}}_{22}) < 0$ *for all* i.
(ii) $\tilde{\mathbf{F}}$, $\tilde{\mathbf{D}}$ *are sufficiently small.*

While the theorem is just what one would expect intuitively, the proof is not trivial because of the nonlinear function $\mathbf{u}^*$. (If $\mathbf{u}^*$ is linear, the proof is an immediate consequence of Theorem 3 and the methods of Example 15).

PROOF. (*a*) Consider first the system (73-*a*), with $\tilde{\mathbf{z}} = \mathbf{0}$. By virtue of (72-A), the following function is clearly positive definite for all $c_0 \geq 0$, $k \geq 0$:

$$V(\mathbf{x}) = \mathbf{x}'\hat{\mathbf{P}}\mathbf{x} - c_0\mathbf{x}'\hat{\mathbf{P}}\hat{\mathbf{D}}\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}\mathbf{x})$$

By simple but somewhat tedious considerations it is verified that

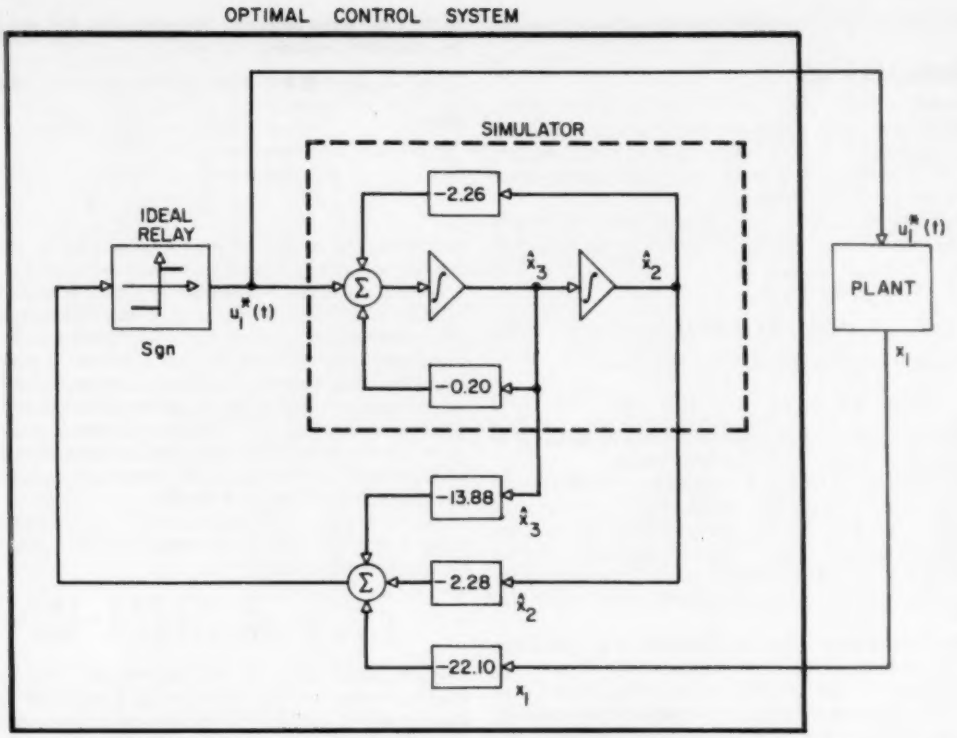


Fig. 14 Design of relay control system

$\dot{V}(\mathbf{x})$ is negative definite, provided that (ii) holds and c_0 is chosen suitably small.

The main steps are as follows. First, note that

$$\dot{V}(\mathbf{x}) \leq -c_1\|\mathbf{x}\|^2 - c_0c_2\|\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}\mathbf{x})\|^2 + (c_0c_3 + c_4)\|\mathbf{x}\| \cdot \|\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}\mathbf{x})\|$$

where

$$c_1 = \min_{\mathbf{x}} \{\mathbf{x}'(\mathbf{Q} - 2\hat{\mathbf{F}}'\hat{\mathbf{P}})\mathbf{x};\ \|\mathbf{x}\| = 1\}$$

$$c_2 = \min_{\mathbf{x}} \{\mathbf{x}'[(\hat{\mathbf{D}} + \bar{\mathbf{D}})'\hat{\mathbf{P}}\bar{\mathbf{D}}]\mathbf{x};\ \|\mathbf{x}\| = 1\}$$

$$c_3 = 2\|\hat{\mathbf{F}}'\hat{\mathbf{P}}\bar{\mathbf{D}}\|$$

$$c_4 = 2\|\bar{\mathbf{P}}\bar{\mathbf{D}}\|$$

Evidently c_1, c_2 can be made positive and c_4 can be made arbitrarily small by virtue of assumption (ii).

$\dot{V}$ will be negative definite, independently of $k > 0$, if

$$c_0c_1c_2 > (c_0c_3 + c_4)^2/4$$

which can be satisfied by taking c_0, c_4 sufficiently small.

(*b*) Next, consider the system (73), with $\bar{\mathbf{F}}_{22} = \mathbf{0}$, in other words, (73-*a*) is coupled to (73-*b*), but not vice versa. Let $\mathbf{M}$ be the matrix which corresponds to $\mathbf{Q} = \mathbf{I}$ and $\hat{\mathbf{F}}_{22}$ through equation (19). By (i), $\mathbf{M}$ will be positive definite.

Now define, as a tentative Lyapunov function for the entire system (73),

$$W(\mathbf{x}, \tilde{\mathbf{z}}) = V(\mathbf{x}) + c_5\|\tilde{\mathbf{z}}\|^2_{\mathbf{M}} \tag{74}$$

In view of part (*a*) and noting that $\|\mathbf{u}^*(\mathbf{x} + \tilde{\mathbf{z}})\| < \|\mathbf{u}^*(\mathbf{x})\| + \|\mathbf{u}^*(\tilde{\mathbf{z}})\|$, the derivative of W can be written as

$$\dot{W}(\mathbf{x}, \tilde{\mathbf{z}}) \leq -c_6\|\mathbf{x}\|^2 - c_7\|\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}\mathbf{x})\|^2 - c_8\|\tilde{\mathbf{z}}\|^2 + [c_8\|\mathbf{x}\| + c_9\|\mathbf{u}^*(k\hat{\mathbf{D}}'\hat{\mathbf{P}}\mathbf{x})\|)]\|\mathbf{u}^*(k\hat{\mathbf{D}}'\bar{\mathbf{P}}\tilde{\mathbf{z}})\|$$

where

$$0 < c_6 < c_1, \qquad 0 < c_7 < c_0c_2$$
$$c_8 = \|\bar{\mathbf{P}}\mathbf{D}'\|, \qquad c_9 = 2\|\mathbf{D}'\bar{\mathbf{P}}\bar{\mathbf{D}}\|$$

Since evidently

$$\|\mathbf{u}^*(k\mathbf{D}'\bar{\mathbf{P}}\tilde{\mathbf{z}})\| < 2c_{10}k\|\tilde{\mathbf{z}}\|$$

it follows that $\dot{W}$ is negative definite if the matrix

$$\begin{bmatrix} -c_6 & 0 & c_9c_{10}k \\ 0 & -c_7 & c_9c_{10}k \\ c_9c_{10}k & c_9c_{10}k & -c_5 \end{bmatrix}$$

is negative definite. This will be true if c_5 is large, of the order of k^2. But the choice of $c_5 > 0$ in (74) is arbitrary. Hence for arbitrarily large, but fixed, k there is a Lyapunov function of type (74) such that

$$\dot{W}(\mathbf{x}, \tilde{\mathbf{z}}) < -c_{11}(\|\mathbf{x}\|^2 + \|\tilde{\mathbf{z}}\|^2)$$

(*c*) Finally, consider the case when $\bar{\mathbf{F}}_{22} \neq \mathbf{0}$. Since by part (*b*), (73) without *this* perturbation is asymptotically stable, with W and W being bounded by *quadratic* forms, hence by the methods of Example 15 it follows that the perturbed system (73) is still asymptotically stable when $\bar{\mathbf{F}}_{22}$ is sufficiently small. Q.E.D.

The magnitude of the permissible errors $\bar{\mathbf{F}}$ and $\bar{\mathbf{D}}$ can be calculated but this is of little interest except in specific cases.

It must be borne in mind also that asymptotic stability of $\hat{\mathbf{F}}$ by no means implies that of $\hat{\mathbf{F}}_{22}$.

Example 18. Design of a Relay Servo. As a numerical illustration of the foregoing design procedure, consider the plant specified by

$$\mathbf{F} = \begin{bmatrix} -0.01 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -2.26 & -0.2 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 0 \\ 0 \\ 2.5 \end{bmatrix}$$

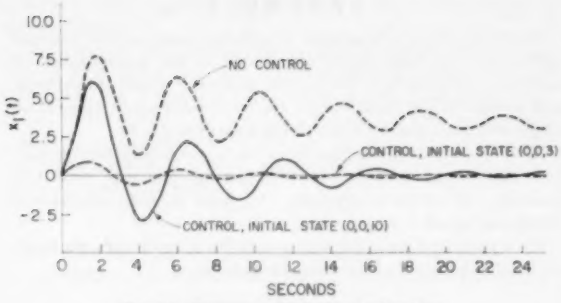


Fig. 15 Response of relay control system

It is easy to see that the transfer function of the plant from $u_1(t)$ to $x_1(t)$ has no zero, a real pole at -0.01 and a pair of complex conjugate poles at $-0.1 \pm i1.5$. Further, assume that $|u_1(t)| \leqq 1$ for all t.

Let $\mathbf{Q} = \mathbf{I}$ and find $\mathbf{P}$ by means of equation (19). Numerical solution leads to

$$\mathbf{P} = \begin{bmatrix} 50.00 & 4.64 & 22.10 \\ 4.64 & 9.71 & 2.28 \\ 22.10 & 2.28 & 13.88 \end{bmatrix}$$

From (72), it follows that the optimal control is given by

$$u_1^*(\mathbf{x}) = \text{sgn}[-22.10x_1 - 2.28x_2 - 13.88x_3]$$

Suppose further that only the output $x_1(t)$ can be measured; the rest of the state variables are to be "synthesized." This leads to the block diagram of Fig. 14. Since $\mathbf{F}_{22}$ is obviously asymptotically stable in this case, it follows that the over-all system shown in the figure is asymptotically stable in the large for small errors in predicting the coefficients of $\mathbf{F}$ and $\mathbf{D}$.

A typical transient response of the uncontrolled plant (i.e., $u_1(t) \equiv 0$ for $t > 0$) is shown in Fig. 15, where $x_1(t)$ is plotted for the initial states $\mathbf{x}_0 = [0, 0, 10]'$, $[0, 0, 3]'$. Note that because of the real pole at -0.01, $x_1(t)$ returns to 0 very slowly. With optimal control and the same initial states $x_1(t)$ goes to zero in about 30 sec even for the largest x_{3_0}—an improvement by about a factor of 10 over the uncontrolled system.

11—Conclusions

The following aspects of the "second method" of Lyapunov are perhaps most significant in the theory of control systems:

(*a*) The "second method" provides an *abstract* tool for studying stability and transient behavior of dynamic systems without solving the differential equations of these systems.

(*b*) Using a Lyapunov function instead of a norm, one has rigorous methods of analysis of nonlinear systems provided that they are (as in most control-system applications) asymptotically stable in the large. This refutes the oft-repeated misconception in the engineering literature that "nonlinear systems cannot be analyzed by exact methods."

(*c*) The abstract method becomes a concrete one whenever explicit expressions can be found for a Lyapunov function. This is always possible in the linear stationary case, by solving a set of simultaneous algebraic equations of order $n(n + 1)/2$. In the linear nonstationary or nonlinear cases, no straightforward methods are available at present for doing this. Further applied mathematical research should be directed toward developing efficient digital computer programs for finding Lyapunov functions.

(*d*) Given a fixed system, the measure of "goodness" of the Lyapunov function is the number Max $V/(-\dot{V})$ (over some region in the state space); this number can be interpreted as the rigorous upper bound on the time constant of the system.

(*e*) Given a system with free parameters, it is natural to identify $-\dot{V}$ with the error criterion to be used for the system. Then V may be identified with the performance index of the system and is to be minimized by proper choice of the free parameters.

(*f*) The optimized system is stable only if the error criterion does not vanish identically along any trajectory. This added condition is necessary to assure the correctness of the usual Wiener theory of optimization.

(*g*) The "second method" of Lyapunov is more a unifying principle than a method. It does not replace ingenuity. On the other hand, in many cases Lyapunov functions are already available, though unrecognized, in standard results in control theory. These results can then be used in other ways to estimate transient response, effect of random perturbations, and so on.

Acknowledgment

It is a pleasure to acknowledge here our indebtedness to Dr. R. W. Bass of RIAS, whose enthusiasm, persistence, and ever-ready help was a most treasured ally against the intricacies of the second method.

References

1 A. M. Lyapunov, "Problème général de la stabilité du mouvement" (in French), *Ann. Fac. Sci. Toulouse*, vol. 9, 1907, pp. 203–474. Reprinted in Ann. Math. Study No. 17, 1949, Princeton University Press.

2 J. L. Massera, "Contributions to Stability Theory," *Ann. Math.*, vol. 64, 1956, pp. 182–206.

3 W. Hahn, "Theorie und Anwendung der Direkten Methode von Lyapunov," (book; in German), *Erg. der Math.*, No. 22, Springer, 1959, 142 pp.

4 I. G. Malkin, "Theory of Stability of Motion" (book), AEC Translation 3352, Dept. of Commerce, USA, 1958.

5 S. Lefschetz, "Differential Equations: Geometric Theory" (book), Interscience, 1957, 364 pp.; see Chapt. 6, par. 2.

6 S. Lefschetz, "Lyapunov and Stability in Dynamical Systems," *Bol. Soc. Mat. Mex.*, vol. 3, 1958, pp. 25–39.

7 A. I. Lur'e, "Some Nonlinear Problems in the Theory of Automatic Control" (book; Russian), Gostekhizdat, Moscow, 1951; German translation Akademie Verlag 1957; English translation, Her Majesty's Stationery Office, 1957.

8 A. M. Letov, "Stability of Nonlinear Regulating Systems" (book; in Russian), *Gosn. Izdat. Tekhn.-Teor. Lit.*, Moscow, 1955.

9 H. Antosiewicz, "A Survey of Lyapunov's Second Method," in "Contr. to Nonlinear Oscillations" (book), IV, Ann. Math. Study 41, 1958, pp. 147–166, Princeton University Press.

10 V. V. Nemitskii and V. V. Stepanov, "Qualitative Theory of Differential Equations," 2nd ed., Moscow, 1951; English translation of 2nd ed., Princeton University Press, 1960; see chapter 5.

11 E. A. Coddington and N. Levinson, "Theory of Ordinary Differential Equations" (book), McGraw-Hill, 1955; see chapter 3.

12 R. E. Kalman and J. E. Bertram, "General Synthesis Procedure for Computer Control of Single and Multi-Loop Linear Systems," *Trans. AIEE*, vol. 77, II, 1958, pp. 602–609.

13 R. E. Kalman and J. E. Bertram, "A Unified Approach to the Theory of Sampling Systems," *J. Franklin Inst.*, vol. 267, 1959, pp. 405–436.

14 R. E. Kalman and R. W. Koepcke, "Optimal Synthesis of Linear Sampling Control Systems Using Generalized Performance Indexes," TRANS. ASME, vol. 80, 1958, pp. 1820–1826.

15 J. E. Bertram and P. E. Sarachik, "Minimum Control Energy System," being presented at the National Automatic Control Conference, Nov. 1959.

16 R. E. Kalman, "Dynamic Optimization of Linear Control Systems. I. Theory," to appear.

17 L. Markus, "Escape Times for Ordinary Differential Equations," *Rend. Sem. Mat.*, Univ. di Torino, vol. 11, 1951–2, pp. 271–276.

18 J. L. Massera, "On Lyapunov's Conditions of Stability," *Ann. Math.* vol. 50, 1949, pp. 705–721.

19 J. E. Bertram and P. E. Sarachik, "Stability of Circuits With Randomly Time-Varying Parameters," Proc. Int. Symp. on Circuit and Information Theory, Los Angeles, Calif., 1959, pp. 260–270.

20 O. Perron, "The Stability Problem in Differential Equations" (in German), *Math. Zeitsch*, vol. 32 1930, pp. 465–473.

21 L. A. Zadeh, "On the Stability of Linear Time-Varying Systems," *J. Appl. Phys.*, vol. 22, 1952, pp. 402–405.

22 I. G. Malkin, "Certain Questions in the Theory of Stability of Motion in the Sense of Lyapunov" (in Russian), *Sbornik Nauchnykh Trudov*, Kazankogo Aviats. Inst., No. 7, 1937; Am. Math. Soc. Translation No. 20, 1950.

23 H. A. Antosiewicz and P. Davis, "Some Implications of Lyapunov's Conditions of Stability," *J. Rat. Math. and Mech.*, vol. 3, 1954, pp. 447–457.

24 R. E. Bellman, "Kronecker Products and the Second Method of Lyapunov," *Math. Nachrichten*, 1959.

25 R. W. Bass, "Routh-Hurwitz Conditions and the Second Method of Lyapunov," *Proc. Am. Soc.*, to appear.

26 H. R. Schwarz, "A Method for Determining Stability of Matrix Differential Equations" (in German), *Z. angew. Math. Phys.*, vol. 7, 1956, pp. 473–500.

27 A. W. Babister, "Determination of the Optimum Response of Linear Systems," *Q. J. Mech. Appl. Math.*, vol. 10, 1957, pp. 360–368, 504–512; vol. 11 ,1958, pp. 119–128.

28 B. S. Razumikhin, "On the Application of Lyapunov's Methods to Stability Problems," *Prikl. Mat. Mekh.*, vol. 22, 1958, pp. 338–349.

29 Ya. N. Roitenberg, "A Method for the Construction of Lyapunov Functions for Linear Systems With Variable Coefficients," *Prikl. Mat. Mekh.*, vol. 22, 1958, pp. 167–172.

30 G. Birkhoff and J. B. Diaz, "Nonlinear Network Problems," *Q. Appl. Math.*, vol. 13, 1955, pp. 431–443; see also R. J. Duffin, "Nonlinear Networks IIa," *Bull. Am. Math. Soc.*, vol. 53, 1947, pp. 963–971.

31 M. A. Aizerman, "On a Problem Concerning the In-the-Large Stability of Dynamic Systems" (in Russian), *Uspekhi Mat. Nauk*, vol. 4, 1949, pp. 187–188.

32 R. E. Kalman, "On Physical and Mathematical Mechanisms of Instability in Nonlinear Automatic Control Systems," TRANS. ASME, vol. 79, 1957, pp. 553–556.

33 E. A. Barbashin, "On the Stability of Solution of a Third-Order Nonlinear Differential Equation," (in Russian), *Prikl. Mat. Mekh.*, vol. 16, 1952, pp. 629–632.

34 M. L. Cartwright, "On the Stability of Solutions of Certain Differential Equations of the Fourth Order," *Q. J. Mech. Appl. Math.*, vol. 9, 1956, pp. 185–194.

35 N. N. Krasovskii, "On the Stability in the Large of a System of Nonlinear Differential Equations" (in Russian), *Prikl. Mat. Mekh.*, vol. 18, 1954, pp. 735–737; also "On Stability Under Large Perturbations" (in Russian), *ibid.*, vol. 21, 1957, pp. 309–319.

36 N. N. Krasovskii, "On the Stability of Solutions of a System of Two Differential Equations" (in Russian), *Prikl. Mat. Mekh.*, vol. 17, 1953, pp. 651–672.

37 G. C. Newton, Jr., L. A. Gould, and J. F. Kaiser, "Analytical Design of Feedback Controls" (book), John Wiley & Sons, New York, N. Y., 1957.

38 R. E. Kalman, "Optimal Control of Saturating Systems by Intermittent Action," IRE WESCON Conv. Record, vol. 1, part 4, 1957, pp. 130–135.

39 C. A. Desoer, "The Bang-Bang Servo Problem Treated by Variational Techniques," *Information and Control*, vol. 2, 1959.

40 J. P. LaSalle, "The Time-Optimal Control Problem," "Contributions to Nonlinear Oscillations" V (book), Annals of Mathematics Studies, Princeton University Press, 1959.

41 R. E. Kalman, "On the General Theory of Control," Moscow Conference on Automatic Control, 1960, to appear.

42 R. Herschel, "On a Generalized Quadratic Optimum," *Regelungstechnik*, vol. 4, 1956, pp. 190–195.

43 J. C. C. McKinsey, "Introduction of the Theory of Games" (book), McGraw-Hill, 1952, 371 pp.

44 R. W. Bass, discussion of a paper by A. M. Letov, Proc. Heidelberg Conf. on Automatic Control ("Regelungstechnik: Moderne Theorien und ihre Verwendbarkeit," by R. Oldenbourg, Munich, 1957), pp. 209–210.

45 I. Flügge-Lotz, "Discontinuous Automatic Control Systems" (book), Princeton University Press, Princeton, N. J., 1953.

46 J. André and P. Seibert, "On Piecewise Linear Differential Equations Occurring in Control Problems" (in German), *Arch. Math.*, vol. 7, 1956, pp. 148–156, 157–164.

47 P. R. Halmos, "Finite-Dimensional Vector Spaces" (book), 2nd Ed., Van Nostrand, 1958.

48 A. S. Householder, "The Approximate Solution of Matrix Problems," *J. Assoc. Comp. Machinery*, vol. 5, 1958, pp. 205–243.

49 R. E. Bellman, "Stability Theory of Differential Equations," McGraw-Hill Book Company, Inc., New York, N. Y., 1953.

APPENDIX

The object here is to present in sufficient detail all the relevant definitions and facts which are needed in the manipulation of norms in applications of the "second method," proofs of theorems, and so on. While these results are very well known, there is no single reference which covers all the necessary points.

The notation at the beginning of Section 4 is used throughout. In calculations, vectors are regarded as matrices with one column. All vectors, matrices, etc. refer to *finite-dimensional* Euclidean spaces.

By a *positive-definite* [*positive semidefinite*] matrix $\mathbf{A}$ we mean any $n \times n$ matrix such that the *quadratic form*

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i,j=1}^{n} x_i a_{ij} x_j$$

is *positive* [*nonnegative*] for all $\mathbf{x} \neq \mathbf{0}$. If $\mathbf{A}$ is positive-definite [semidefinite] then $-\mathbf{A}$ is *negative-definite* [*semidefinite*]. It is not required as a rule that $\mathbf{A}$ be *symmetric* ($\mathbf{A} = \mathbf{A}'$) but note that if $\mathbf{B}$ is any *antisymmetric* matrix $\mathbf{B} = -\mathbf{B}'$ then $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'(\mathbf{A} + \mathbf{B})\mathbf{x}$.

The *eigenvalues* $\lambda_i(\mathbf{A})$ of the $n \times n$ matrix $\mathbf{A}$ are the n-complex roots of the polynomial $\det(\mathbf{A} - \lambda\mathbf{I})$. A positive-definite matrix $\mathbf{P}$ always has an inverse, in fact $\lambda_i(\mathbf{P}) > 0$ for all i. Note also that $\lambda_i(\mathbf{AB}) = \lambda_i(\mathbf{BA})$ and $\lambda_i(\mathbf{A}^{-1}) = \lambda_i^{-1}(\mathbf{A})$.

It can be shown that *a matrix* $\mathbf{A}$ *is positive-definite* [*semidefinite*] *if and only if any of the following conditions hold:*

(i) *There is a nonsingular* [*singular*] *matrix* $\mathbf{B}$ *such that* $\mathbf{B}'\mathbf{B} = \mathbf{A}$;

(ii) $\lambda_i(\mathbf{A}) > 0$ [$\lambda_i(\mathbf{A}) \geqq 0$] *for all* i*:*

(iii) *all principal minors of* $\mathbf{A}$ *are positive* [*nonnegative*].

The concept of the norm is similar to that of the absolute value. It is needed in order to be able to make statements about convergence, continuity, etc. in Euclidean space. Standard references are: Halmos [47], chapter 3; Householder [48].

Specifically, a *norm* is a function which assigns to every vector (or point) $\mathbf{x}$ in a given Euclidean space a real number $\|\mathbf{x}\|$ such that

(N_1) $\|\mathbf{x}\| \geqq 0$ for all $\mathbf{x}$

(N_2) $\|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}$, $\mathbf{y}$

(N_3) $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for all $\mathbf{x}$ and complex constant α

(N_4) $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = \mathbf{0}$

If axiom (N_4) is missing, then $\|\mathbf{x}\|$ is a *seminorm*. Examples:

(a) $\|\mathbf{x}\| = \sum_{i=1}^{n} |x_i|$

(b) $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} = (\mathbf{x}'\mathbf{x})^{1/2}$

This is the standard definition of the distance of a point in a Euclidean space from the origin, also called *Euclidean norm.*

(c) $\|\mathbf{x}\|_{\mathbf{A}} = (\mathbf{x}'\mathbf{A}\mathbf{x})^{1/2}$ where $\mathbf{A}$ is symmetric and positive definite; called here the *generalized Euclidean norm.*

(d) $\|\mathbf{x}\| = \operatorname{Max}_i \{|x_i|\}$

In explicit calculations, $\|\mathbf{x}\|$ always means the Euclidean norm.

If $\mathbf{A}$ is positive semidefinite then $\|\mathbf{x}\|_{\mathbf{A}}$ is a *seminorm.*

By a *topology* on a mathematical space $\mathfrak{X}$ we mean a collection of subsets of $\mathfrak{X}$ (called *open sets*) satisfying certain axioms which serve to define abstractly the concept of nearness between elements of $\mathfrak{X}$. For a Euclidean space, a topology is usually defined as follows: Call the set of all points $\|\mathbf{x} - \mathbf{y}\| < r$ an *open r-sphere* about $\mathbf{y}$. The topology then consists of all r-spheres ($0 < r < \infty$) about every point $\mathbf{x}$ and their unions and finite intersections. A

Euclidean space together with a given norm is then called a *normed space*.

Convergence and continuity in a space is always defined with respect to a given topology. Two topologies are *equivalent* if every open set of one topology contains an open set and is contained in an open set of the other topology. Two norms in a given finite dimensional Euclidean space define equivalent topologies since it is always possible to find two positive constants c_1, c_2 such that [48]

$$c_1\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq c_2\|\mathbf{x}\|_1 \quad \text{for all } \mathbf{x}^{14}$$

For example if $\|\mathbf{x}\|_1$ is defined as in (a) above and $\|\mathbf{x}\|_2$ as in (*b*) above, then it follows from elementary calculations (ref. [11], p. 17)

$$n^{-1/2}\sum_{i=1}^{n}|x_i| \leq (\mathbf{x}'\mathbf{x})^{1/2} \leq \sum_{i=1}^{n}|x_i| \tag{75}$$

The method used in Example 14 proves also the following relations between generalized Euclidean norms:

$$\lambda_{\min}{}^{1/2}(\mathbf{Q}\mathbf{P}^{-1})\|\mathbf{x}\|_{\mathbf{P}} \leq \|\mathbf{x}\|_{\mathbf{Q}} \leq \lambda_{\max}{}^{1/2}(\mathbf{Q}\mathbf{P}^{-1})\|\mathbf{x}\|_{\mathbf{P}} \tag{76}$$

($\mathbf{Q}$, $\mathbf{P}$ symmetric and positive-definite).

Closely related to the concept of the norm of a vector is the *norm of a matrix*. This is defined by

$$\|\mathbf{A}\| = \text{Min } K \text{ such that } \|\mathbf{A}\mathbf{x}\| \leq K\|\mathbf{x}\|$$

For the Euclidean norm, this definition is obviously equivalent to:

$$\|\mathbf{A}\|^2 = \max_{\mathbf{x}}\{\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x};\ \mathbf{x}'\mathbf{x} = 1\}$$

[14] Therefore, in questions of convergence, stability, etc., it is immaterial as to what norm is used.

The following inequalities are obvious from the definition

$$\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\| \tag{77}$$

$$\|\mathbf{A}'\| = \|\mathbf{A}\| \tag{78}$$

An explicit formula for the Euclidean norm follows at once from the second definition and formula (76):

$$\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A}) \tag{79}$$

Since $\lambda_i(\mathbf{A}^2) = \lambda_i{}^2(\mathbf{A})$, for a *symmetric* matrix $\mathbf{A}$ the expression for the Euclidean norm is

$$\|\mathbf{A}\| = \lambda_{\max}{}^{1/2}(\mathbf{A}'\mathbf{A}) = \lambda_{\max}{}^{1/2}(\mathbf{A}^2) = |\lambda(\mathbf{A})|_{\max} \tag{80}$$

In other cases explicit expressions for the norm are usually not available.

One can easily compute some (more or less pessimistic) upper bounds corresponding to the various norms defined above

$$\|\mathbf{A}\|_a \leq \sum_{j,i=1}^{n}|a_{ij}|$$

$$\|\mathbf{A}\|_b \leq \left(\sum_{j,i=1}^{n} a_{ij}{}^2\right)^{1/2}$$

$$\|\mathbf{A}\|_c \leq \max_i\left\{\sum_{j=1}^{n}|a_{ij}|\right\}$$

Control System Analysis and Design Via the "Second Method" of Lyapunov[1]

II Discrete-Time Systems

R. E. KALMAN
Research Institute for Advanced Study,[2] Baltimore, Md.

J. E. BERTRAM
IBM Research Center, Yorktown Heights, N. Y.

The second method of Lyapunov is applied to the study of discrete-time (sampled-data) systems. With minor variations, the discussion parallels that of the companion paper on continuous-time systems. Theorems are stated in full but motivation, proofs, examples, and so on, are given only when they differ materially from their counterparts in the continuous-time case.

1—Introduction

We apply here the considerations of the companion paper,[3] "Control System Analysis and Design Via the 'Second Method' of Lyapunov. I Continuous-Time Systems," to the study of dynamic systems in which the time variable changes discretely, i.e., those governed by ordinary *difference* (rather than differential) equations.

In order to avoid needless duplication, all equations, theorems, examples marked by an asterisk will refer to corresponding items of the companion paper. Other equations, etc., will be numbered consecutively. All theorems will be stated in full, but examples, proofs, etc., will be given only if they differ in a nontrivial way from their continuous-time counterparts. A few results peculiar to the discrete-time case are included, however, with full discussion.

The need for a separate presentation is due to the increasing importance of sampled-data systems in connection with digital-computer control. The development of the study of continuous-time and discrete-time systems has tended to drift apart somewhat as a result of the popularity of the z-transform approach to the latter [1].[4] We shall endeavor to emphasize here the essential unity as well as some of the peculiar contrasts of the two subjects, from a point of view rather different from the z-transform.

2—Outline of Contents

In Section 4 we present a method of description of discrete-time dynamic systems which is analogous to the terminology now current in mathematics for the description of continuous-time dynamic systems. The important concepts are those of the state and state transition—these notions are indispensable not only for the application of the "second method" but even for a rigorous definition of stability. Section 5 recalls the customary definitions of stability; these are fully identical with the ones used in the continuous-time case.

Section 6 presents the main theorems on which the "second method" is based. Again, the analogy with the continuous-time case is complete and the reader is referred to the companion paper[3] for motivation and proofs.

Section 7 deals with applications to stability theory. Example 6* presents the derivation of the Routh-Hurwitz inequalities via the "second method." Since these inequalities are not well known in the discrete-time case, a full derivation is given. Theorem 4* and Example 2 illustrate the concept of a *contraction;* this simple notion leads to useful nonlinear stability criteria which are at present almost unknown in the sampled-data systems literature.

Sections 8, 9, and 10 are analogous with those of the companion paper, with minor variations in the details of the derivations.

3—Guide to Western Literature

There is very little literature on the stability theory of systems governed by difference equations, since such problems are usually trivial special cases of the stability theory of continuous-time systems. An early paper is that of Ta Li [2]. Recently, Hahn [3] has provided a self-contained summary of the application of the "second method" of Lyapunov to difference equations; this material is also included in his book [4].

An important prerequisite to the application of the "second method" is that system equations must be given from the "state" point of view. Often these equations cannot be obtained in a *natural* and convenient way from the z-transform [1] type of description of discrete-time systems, particularly when the system contains dynamic elements of two types—those governed by difference or by differential equations. A detailed treatment of the derivation of the discrete-time dynamic equations in cases of current practical interest has been given recently by the authors [5]. This paper provides a complete treatment of the linear case, including the analysis of sampling operations of complicated type (see also Reference [6]).

4—Description of Discrete-Time Dynamic Systems

Small boldface Roman or Greek letters will denote vectors. Capital boldface Roman or Greek letters will denote matrices. Unless otherwise specified, all scalars and all elements of vectors and matrices will be real numbers. The prime denotes the transpose of a vector or matrix; thus the inner (scalar) product is written as $\mathbf{x}'\mathbf{y}$ and a quadratic form as $\mathbf{x}'\mathbf{A}\mathbf{x}$. The norm is denoted by $\|\mathbf{x}\|$; in specific calculations this means always the Euclidean norm $(\mathbf{x}'\mathbf{x})^{1/2}$. The eigenvalues of a matrix $\mathbf{F}$ are $\lambda_i(\mathbf{F})$. The unit matrix is $\mathbf{I}$.

---

[1] This research was supported in part by the U. S. Air Force Office of Scientific Research under Contract AF 49(638)-382.

[2] 7212 Bellona Ave.

[3] R. E. Kalman and J. E. Bertram, "Control System Analysis and Design Via the 'Second Method' of Lyapunov. I Continuous-Time Systems," published in this issue, pp. 371–393.

[4] Numbers in brackets designate References at end of paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE-IRE-ISA National Automatic Control Conference, Dallas, Tex., November 4–5, 1959.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 24, 1959. Paper No. 59—NAC-3.

In this paper, we shall study systems governed by the *vector difference* equation.

$$\mathbf{x}(t_{k+1}) = \mathbf{h}(\mathbf{x}(t_k), \mathbf{u}(t_k), t_k) \qquad (5^*)$$

where the t_k indicate discrete values of time (k = integer)

$$-\infty < \ldots, t_{k-1} < t_k < t_{k+1}, \ldots < +\infty; \quad t_k \to \infty \text{ when } k \to \infty$$

at which the behavior of the system can be or is observed; t_k is regarded as an independent variable analogous to t in the continuous-time case. Of course, (5*) is equivalent to the set of n scalar difference equations

$$x_i(t_{k+1}) = h_i(x_1(t_k), \ldots, x_n(t_k), u_1(t_k), \ldots, u_m(t_k), t_k) \quad (5^*)$$

$$i = 1, \ldots, n$$

The vector $\mathbf{x}$ is the *state* of the system (5*), its components x_i are the *state variables*. The vector $\mathbf{u}(t_k)$ is the *control* function (or *forcing function* or *input*) of the system (5*); its components $u_i(t)$ are the *control variables*. The system is specified by the vector-valued function $\mathbf{h}$. It is always assumed that $\mathbf{h}$ is a 1-to-1 function for any fixed $\mathbf{u}$ and t_k, continuous in all of its arguments. The integer n is the *order* of the system. Usually, $m < n$. To assure that (5*) represents a physical system, we must require also that if $t_{k+1} - t_k \to 0$ as $k \to \infty$, then $\mathbf{h}(\mathbf{x}, \mathbf{u}, t_k) \to \mathbf{x}$.

If $\mathbf{u}(t_k) \equiv \mathbf{0}$ for all t_k, we say that (5*) is *free* (*unforced*):

$$\mathbf{x}(t_{k+1}) = \mathbf{h}(\mathbf{x}(t_k), t_k) \qquad (5^*\text{-F})$$

Equation (5*) defines a *discrete-time dynamic system*. Unlike in the case of continuous-time dynamic systems which are usually specified by their infinitesimal state-transitions (i.e., by differential equations) there is no difficulty (i.e., no finite escape time) here to obtain from the primary description (5*) *all* possible state transitions.

For any initial state $\mathbf{x}_0$, any initial time t_0, and any time t_k, we define by induction the function

$$\boldsymbol{\phi}(t_k;\ \mathbf{x}_0, t_0) = \mathbf{x}(t_k)$$

satisfying $\quad \boldsymbol{\phi}(t_0;\ \mathbf{x}_0, t_0) = \mathbf{x}_0 \quad \text{for all} \quad \mathbf{x}_0, t_0 \qquad (7^*)$

$$\boldsymbol{\phi}(t_{k+1};\ \mathbf{x}(t_k), t_k) = \mathbf{h}(\mathbf{x}(t_k), \mathbf{u}, t_k) \quad \text{for all} \quad \mathbf{x}(t_k),\ t_k,$$

and some *fixed* sequence of values $\mathbf{u}(t_0), \ldots, \mathbf{u}(t_k), \ldots$ of the forcing function

We call $\boldsymbol{\phi}(t_k;\ \mathbf{x}_0, t_0)$ (where t_k takes on all possible discrete values) the *motion* (or *trajectory*) of the system (5*) going through state $\mathbf{x}_0$ at time t_0. Since the function $\mathbf{h}$ in (5*) is unique and one-to-one, it follows that for any $\mathbf{x}_0$ and any discrete values t_c, t_b, t_a of time

$$\boldsymbol{\phi}(t_c;\ \mathbf{x}_0, t_a) = \boldsymbol{\phi}(t_c;\ \boldsymbol{\phi}(t_b;\ \mathbf{x}_0, t_a), t_b) \qquad (8^*)$$

If $\mathbf{h}$ is not one-to-one, (8*) holds if $t_a \leqq t_b \leqq t_c$. It is usually not necessary to indicate explicitly the dependence of the motion on the forcing function.

A state $\mathbf{x}_e$ of a free dynamic system (5*-F) is an *equilibrium state* if

$$\mathbf{h}(\mathbf{x}_e, t_k) = \mathbf{x}_e \quad \text{for all} \quad t_k \qquad (10^*)$$

A dynamic system (5*) is *stationary* if $\mathbf{h}(\mathbf{x}, \mathbf{u}, t_k) \equiv \mathbf{h}(\mathbf{x}, \mathbf{u})$, i.e., $\mathbf{h}$ is not an explicit function of t_k. Then we take for simplicity $t_{k+1} - t_k$ = constant for all k. If a system is both free and stationary, it is *autonomous*.

A dynamic system (5*) is *linear* if $\mathbf{h}$ is a linear function of $\mathbf{x}$ and $\mathbf{u}$.

If $\mathbf{x}_e = \mathbf{0}$ is an equilibrium state and if (5*) is linear, then we can write

$$\mathbf{x}(t_{k+1}) = \mathbf{H}(t_k)\mathbf{x}(t_k) + \boldsymbol{\Delta}(t_k)\mathbf{u}(t_k) \qquad (5^*\text{-L})$$

where

$$\mathbf{H}(t_k) = [\partial h_i(t_k)/\partial x_j]$$

$$\boldsymbol{\Delta}(t_k) = [\partial h_i(t_k)/\partial u_j]$$

Let us define

$$\boldsymbol{\Phi}(t_k, t_k) = \mathbf{I} \quad \text{for all} \quad t_k$$
$$\boldsymbol{\Phi}(t_{k+1}, t_l) = \mathbf{H}(t_k) \ldots \mathbf{H}(t_l) \quad \text{for all} \quad t_k \geqq t_l \qquad (7^*\text{-L})$$

As in the continuous case, the matrix $\boldsymbol{\Phi}(t_k, t_0)$ is called the *transition matrix* of (5*-L); but now its arguments t_k, t_0 are defined for discrete values of time only.

Now we can write a general expression for motions of (5*-L), valid for all $t_k \geqq t_0$

$$\boldsymbol{\phi}(t_k;\ \mathbf{x}_0, t_0) = \boldsymbol{\Phi}(t_k, t_0)\mathbf{x}_0 + \sum_{t_0 \leqq t_i < t_k} \boldsymbol{\Phi}(t_k, t_i)\boldsymbol{\Delta}(t_i)\mathbf{u}(t_i) \qquad (6^*\text{-L})$$

If $\mathbf{H}(t_k)$ is nonsingular for every t_k, then its inverse is

$$\mathbf{H}^{-1}(t_k) = \boldsymbol{\Phi}^{-1}(t_{k+1}, t_k) = \boldsymbol{\Phi}(t_k, t_{k+1})$$

Then the transition matrix $\boldsymbol{\Phi}(t_k, t_l)$ is defined also for $t_l > t_k$ and we can write for any discrete t_a, t_b, t_c

$$\boldsymbol{\Phi}(t_c, t_a) = \boldsymbol{\Phi}(t_c, t_b)\boldsymbol{\Phi}(t_b, t_a) \qquad (8^*\text{-L})$$

and also

$$\boldsymbol{\Phi}^{-1}(t_a, t_b) = \boldsymbol{\Phi}(t_b, t_a) \qquad (9^*)$$

If $\mathbf{H}(t_k)$ is singular, then (8*-L) holds if we require that $t_a \leqq t_b \leqq t_c$.

If the linear system (5*-L) is stationary, then $\mathbf{H}(t_k)$, $\boldsymbol{\Delta}(t_k)$ are constants and $\boldsymbol{\Phi}(t_k, t_0)$ depends only on the difference $t_k - t_0$.

Example 1—Discretization of Continuous-Time Systems. A discrete-time dynamic system may arise from a continuous-time dynamic system by simply considering the latter at discrete instants of time only. Then our present notation coincides exactly with that of the companion paper.[3]

But not every discrete-time system can be regarded arising in a *natural* way from a continuous-time system. For instance, if $\boldsymbol{\Phi}(t_k, t_0)$ is the transition matrix of a linear, stationary, discrete-time system, obtained from a linear, stationary, continuous-time system, then we would have

$$\boldsymbol{\Phi}(t_k, t_0) \equiv \boldsymbol{\Phi}(t_k - t_0) = \exp\,(t_k - t_0)\mathbf{F} \qquad (1)$$

But it is well known that the foregoing equation cannot be satisfied with any real, constant matrix $\mathbf{F}$ if $\boldsymbol{\Phi}(t_k - t_0)$ has any real, nonpositive eigenvalues. A simple instance of this is the equation

$$x(t_{k+1}) = -x(t_k)/2$$

Note also that the discrete version of a continuous-time stationary system is stationary only if $t_{k+1} - t_k$ = constant for all t_k.

Example 2—Sampling Systems. A *sampling* (or *sampled-data*) system is one in which some variables are allowed to change only at discrete instants t_k of time. These instants may specify the time at which some physical measurement (say, radar pulse) is performed, the time at which the memory of a digital computer is read out, etc. There are several different types of sampling operations of practical significance:

(*a*) *Periodic* (*conventional*) *sampling:* $t_k = kT$, where T = constant > 0.

(*b*) *Multiple-order sampling:* The pattern of t_k's is repeated periodically, i.e., $t_{k+r} - t_k$ = constant for all k.

(*c*) *Multirate sampling:* In case of two concurrent sampling operations, $t_k = pT_1$ or qT_2 where T_1, T_2 are constants and p, q are integers.

(*d*) *Random sampling:* t_k is a random variable.

In case (*a*), the discrete dynamic system is stationary; in the

remaining cases the system is nonstationary in general. In case (b), the system can be made stationary by considering only sampling points t_k, t_{k+r}, t_{k+2r}, . . . In case (c), the system can be made stationary by taking samples $L(T_1, T_2)$ units of time apart where $L(T_1, T_2)$ is the least common multiple of T_1, T_2; this fails, of course, if $L = \infty$; i.e., if T_1/T_2 is irrational.

The "second method" of Lyapunov, as usually stated, applies only to free dynamic systems with equilibrium state at $\mathbf{0}$. More generally, this means that the method is concerned only with deviations about some fixed motion (see the companion paper[2] for further details).

5—Concepts of Stability

The definitions of this section, as well as their motivation, are identical in every respect with those of the companion paper[2] except that the concept of finite escape time is missing.

(S_1) An equilibrium state $\mathbf{x}_e$ of a dynamic system is *stable* [(S_6): *uniformly stable*] if to any number $\epsilon > 0$ there corresponds a number $\delta(\epsilon, t_0) > 0[\delta(\epsilon) > 0]$ such that if $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq \delta$ then $\|\phi(t_k; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq \epsilon$ for all $t_k \geq t_0$.

(S_2) An equilibrium state $\mathbf{x}_e$ of a dynamic system is *equiasymptotically stable* [(S_7): *uniformly asymptotically stable*] if

(a) it is stable [uniformly stable] and

(b) to any number $\mu > 0$ there corresponds a number $T(\mu, t_0)$ $[T(\mu)]$ such that $\|\phi(t_k; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leq \mu$ for all $t_k \geq t_0 + T$ whenever $\|\mathbf{x}_0 - \mathbf{x}_e\| \leq r$; $r(t_0) > 0[r > 0]$ being some fixed constant which does not depend on μ or $\mathbf{x}_0$.

(S_4) An equilibrium state $\mathbf{x}_e$ of a dynamic system is *equiasymptotically stable in the large* [(S_8): *uniformly asymptotically stable in the large*] if

(a) it is stable [uniformly stable];

(b) all motions are bounded [uniformly bounded];

(c) all motions $\phi(t_k; \mathbf{x}_0, t_0)$ ($\mathbf{x}_0$, t_0 arbitrary) converge [converge uniformly in $\|\mathbf{x}_0\| \leq r$, t_0; r being arbitrarily large] to $\mathbf{x}_e$ with increasing t_k.

The various implications between these definitions, as stated in the companion paper,[2] hold in the discrete case also.

6—The Main Theorems

We state here results corresponding to those of the companion paper.[2] After obvious modifications (replacing continuous t with discrete t_k, integrals by sums, etc.) the proofs are identical with those given in the companion paper and are therefore omitted (see Hahn [3] for a self-contained account).

Theorem 1*. *Consider the discrete-time, free dynamic system:*

$$\mathbf{x}(t_{k+1}) = \mathbf{h}(\mathbf{x}(t_k), t_k) \qquad (5^*\text{-F})$$

where $\mathbf{h}(\mathbf{0}, t_k) = \mathbf{0}$ *for all* t_k. SUPPOSE *there exists a scalar function* $V(\mathbf{x}, t_k)$ *such that* $V(\mathbf{0}, t_k) = 0$ *for all* t_k, *and*

(i) $V(\mathbf{x}, t_k)$ *is positive definite; i.e., there exists a continuous, nondecreasing scalar function* α *such that* $\alpha(0) = 0$ *and, for all* t_k *and all* $\mathbf{x} \neq \mathbf{0}$,

$$0 < \alpha(\|\mathbf{x}\|) \leq V(\mathbf{x}, t_k);$$

(ii) *There exists a continuous scalar function* γ *such that* $\gamma(0) = 0$ *and, for all* t_k *and all* $\mathbf{x} \neq \mathbf{0}$,

$$\begin{aligned}\Delta V(\mathbf{x}, t_k) &= \text{rate of increase of } V \text{ along motion starting at } \mathbf{x}, t_k \\ &= [V(\phi(t_{k+1}; \mathbf{x}, t_k), t_{k+1}) - V(\mathbf{x}, t_k)](t_{k+1} - t_k)^{-1} \\ &\leq -\gamma(\|\mathbf{x}\|) < 0^5;\end{aligned}$$

(iii) *There exists a continuous, nondecreasing scalar function* β *such that* $\beta(0) = 0$ *and, for all* t_k *and all* $\mathbf{x} \neq \mathbf{0}$,

$$V(\mathbf{x}, t_k) \leq \beta(\|\mathbf{x}\|);$$

(iv) $\alpha(\|\mathbf{x}\|) \to \infty$ *when* $\|\mathbf{x}\| \to \infty$.

THEN *the equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *is uniformly asymptotically stable in the large and* $V(\mathbf{x}, t_k)$ *is a Lyapunov function of the system* (5*-F).

Corollary 1.1*. *The following conditions are sufficient for various weaker types of stability:*

1 *Uniform asymptotic stability:* (i–iii).
2 *Equiasymptotic stability in the large:* (i–ii), (iv).
3 *Equiasymptotic stability:* (i–ii).
4 *Uniform stability:* (i), (iii), and (ii_1): $\Delta V(\mathbf{x}, t_k) \leq 0$ for all $\mathbf{x}$, t_k.
5 *Stability:* (i–ii_1).

Corollary 1.2*. *If the system* (5*) *is autonomous, it suffices to take* V *as a function of* $\mathbf{x}$ *only, such that* $V(\mathbf{0}) = 0$ *and*

(i-A) $V(\mathbf{x}) > 0$ *when* $\mathbf{x} \neq \mathbf{0}$;
(ii-A) $\Delta V(\mathbf{x}) < 0$ *when* $\mathbf{x} \neq \mathbf{0}$;
(iii-A) $V(\mathbf{x})$ *is continuous in* $\mathbf{x}$;
(iv-A) $V(\mathbf{x}) \to \infty$ *when* $\|\mathbf{x}\| \to \infty$.

Corollary 1.3*. *In Corollary 1.2*, Condition* (ii-A) *may be replaced by*

(ii_1-A) $\Delta V(\mathbf{x}) \leq 0$ *for all* $\mathbf{x}$;
(ii_2-A) $\Delta V(\phi(t_k; \mathbf{x}_0, t_0))$ *does not vanish identically in* $t_k \geq t_0$ *for any* t_0 *and any* $\mathbf{x}_0 \neq \mathbf{0}$.

The problem of the converse of Theorem 1* (i.e., does uniform asymptotic stability in the large imply the existence of a Lyapunov function with properties (i-iv)?) apparently has not been investigated so far. Theorem 2 of the companion paper[2] does not carry over trivially in this case because, as we have seen in Example 1, the class of all discrete-time dynamic systems is *larger* than the class of all continuous-time dynamic systems.

Theorem 3*. *Consider the discrete-time, linear, dynamic system*

$$\mathbf{x}(t_{k+1}) = \mathbf{H}(t_k)\mathbf{x}(t_k) + \mathbf{\Delta}(t_k)\mathbf{u}(t_k) \qquad (5^*\text{-L})$$

and ASSUME *that*

(i) $(t_k - t_{k-1})^{-1} \cdot \|\mathbf{H}(t_k) - \mathbf{1}\| \leq c_1 < \infty$ *for all* t_k[6]

(ii) $0 < c_2 \leq (t_{k+1} - t_k)^{-1} \cdot \|\mathbf{\Delta}(t_k)\mathbf{x}\| \leq c_3 < \infty$ *for all* $\|\mathbf{x}\| = 1$, *all* t_k

THEN *the following propositions concerning this system are equivalent:*

(A) *Assuming* $\mathbf{x}_0 = \mathbf{0}$, *any uniformly bounded excitation*

$$\|\mathbf{u}(t_k)\| \leq c_4 < \infty \text{ when } t_k \geq t_0$$

gives rise to a uniformly bounded response for all $t_k \geq t_0$, i.e.

$$\begin{aligned}\|\mathbf{x}(t_k)\| &= \Big\| \sum_{t_0 \leq t_i < t_k} \Phi(t_k, t_{i+1})\mathbf{\Delta}(t_i)\mathbf{u}(t_i) \Big\| \\ &\leq c_5(c_4) < \infty;\end{aligned}$$

(B) *For all* $t_k > t_0$,

$$\sum_{t_0 < t_i \leq t_k} (t_i - t_{i-1}) \|\Phi(t_k, t_i)\| \leq c_6 < \infty;$$

(C) *The equilibrium state* $\mathbf{x}_e = \mathbf{0}$ *of the free system is uniformly asymptotically stable;*

(D) *There exist positive constants* c_7, c_8 *such that, whenever* $t_k \geq t_0$,

[5] When the discrete-time system is given abstractly (i.e., not by discretization of a continuous-time system), there is no loss of generality in assuming that $t_{k+1} - t_k = 1$ for all k.

[6] In this theorem, c_1, c_2, . . . denote various positive constants.

$$\|\Phi(t_k, t_0)\| \leq c_9 e^{-c_8(t_k - t_0)}; \qquad (15^*)$$

(*E*) *Given any positive definite matrix* $\mathbf{Q}(t_k)$ *satisfying, for all* $t_k \geq t_0$,

$$0 < c_9 \mathbf{I} \leq \mathbf{Q}(t_k) \leq c_{10} \mathbf{I} < \infty^7$$

the scalar function defined by

$$V(\mathbf{x}, t_k) = \sum_{t_k \leq t_i \leq \infty} (t_{i+1} - t_i) \|\Phi(t_i, t_k)\mathbf{x}\|^2_{\mathbf{Q}(t_i)}{}^8 \qquad (17^*)$$

$$= \|\mathbf{x}\|^2_{\mathbf{P}(t_k)} \qquad (18^*)$$

exists[9] *and is a Lyapunov function for the free system, satisfying the requirements of Theorem 1* with*

$$\Delta V(\mathbf{x}, t_k) = -\|\mathbf{x}\|^2_{\mathbf{Q}(t_k)}$$

Corollary 3.1*. *A discrete-time, free, linear, stationary, dynamic system*

$$\mathbf{x}(t_{k+1}) = \mathbf{H}\mathbf{x}(t_k) \qquad (t_{k+1} - t_k = 1) \qquad (5^*\text{-FLS})$$

is asymptotically stable if, and only if, given any symmetric, positive-definite matrix $\mathbf{Q}$ *there exists a symmetric, positive-definite matrix* $\mathbf{P}$ *which is the unique solution of the linear equation*

$$\mathbf{H}'\mathbf{P}\mathbf{H} - \mathbf{P} = -\mathbf{Q} \qquad (19^*)$$

and

$$V(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{P}}$$

is a Lyapunov function for (5-FLS) with*

$$\Delta V(\mathbf{x}) = -\|\mathbf{x}\|^2_{\mathbf{Q}}$$

Corollary 3.2.* *The eigenvalues of a constant matrix* $\mathbf{H}$ *are less than* ρ *in absolute value if, and only if, given any symmetric, positive definite matrix* $\mathbf{Q}$ *the linear equation*

$$\rho^{-2}\mathbf{H}'\mathbf{P}\mathbf{H} - \mathbf{P} = -\mathbf{Q} \qquad (20^*)$$

has a unique, symmetric, positive-definite solution $\mathbf{P}$.

The proof of Theorem 3 and its corollaries is entirely analogous to the proof in the continuous-time case.

7—Applications to Stability Theory

7.1 Routh-Hurwitz Conditions for Linear Stationary Systems. As in the companion paper,[3] Corollary 3.1* gives a purely algebraic procedure for checking the asymptotic stability of a free, linear, stationary system. (As is well known, a necessary and sufficient condition for this is that all eigenvalues of the matrix $\mathbf{H}$ in (5* FLS) be less than unity in absolute value.) Take any symmetric, positive-definite $\mathbf{Q}$ (say, the unit matrix); invert (19*) to find $\mathbf{P}$, which is always possible in case of asymptotic stability; test $\mathbf{P}$ for positive-definiteness, in other words, require the leading principal minors of $\mathbf{P}$ to be positive—these n inequalities are analogous to the Routh-Hurwitz conditions.

This procedure is illustrated in Example 6*. The procedure is most useful for machine computation when the number of nonzero elements of $\mathbf{H}$ is larger than about 4 or 5; in such cases, a stability test may require very extensive numerical computations by conventional methods. The analog of the canonic matrix (33) (Example 7) of the companion paper is so far not yet available for discrete systems.

Example 6*. Routh-Hurwitz Conditions for General Second-Order Case. Consider the case when $\mathbf{H}$ is a 2×2 matrix. Direct substitution of the general expression for $\mathbf{H}$ into (19*) leads to very complicated algebra so that some simplification must be sought.

If $\mathbf{H} = \lambda\mathbf{I}$, then $\lambda^2 = \det \mathbf{H} < 1$ insures stability. If $\mathbf{H} \neq \lambda\mathbf{I}$, then it is easy to show that there is some nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{H}\mathbf{T} = \mathbf{C} = \begin{bmatrix} 0 & 1 \\ -a_1 & -a_2 \end{bmatrix}$$

Writing out (19*) in detail for $\mathbf{H} = \mathbf{C}$, with $\mathbf{Q} = \mathbf{I}$, we get

$$\begin{bmatrix} -1 & 0 & a_1^2 \\ 0 & -1 - a_1 & a_1 a_2 \\ 1 & -2a_2 & a_2^2 - 1 \end{bmatrix} \cdot \begin{bmatrix} p_{11} \\ p_{12} \\ p_{22} \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}$$

This set of equations has a unique solution if the determinant of the matrix on the left does not vanish:

$$\det \mathcal{C} = (a_1 - 1)[(1 + a_1)^2 - a_2^2] \neq 0 \qquad (26^*)$$

Assuming (26*), we find after elementary computations

$$\mathbf{P} = (\det \mathcal{C})^{-1} \begin{bmatrix} \det \mathcal{C} - 2a_1^2(1 + a_1) & -2a_1 a_2 \\ -2a_1 a_2 & -2(1 + a_1) \end{bmatrix} \qquad (27^*)$$

Now $\mathbf{P}$ is positive definite if, and only if,

$$\det \mathbf{P} = \frac{2(1 + a_1^2)}{(1 - a_1)^2[(1 + a_1)^2 - a_2^2]} > 0 \qquad (28^*)$$

$$p_{22} = \frac{2(1 + a_1)}{(1 - a_1)[(1 + a_1)^2 - a_2^2]} > 0 \qquad (29^*)$$

But (28*) implies that

$$(1 + a_1)^2 - a_2^2 > 0$$

and this fact, with (29*) implies that

$$1 - a_1^2 > 0$$

Now observe that $a_1 = \det \mathbf{C}$ and $a_2 = -\operatorname{tr} \mathbf{C}$. Since the determinant and the trace are invariant under similarity transformations, it follows that we can write equally well

$$|\operatorname{tr} \mathbf{H}| < |1 + \det \mathbf{H}| \qquad (30^*)$$

$$|\det \mathbf{H}| < 1 \qquad (31^*)$$

which is the general form of the Routh-Hurwitz inequalities in the second-order case. Finally, it follows from (30*–31*) that condition (26*) is always satisfied.

7.2 Stability in Nonlinear Systems. The analog of Krasovskii's theorem (Theorem 4) of the companion paper[3] is well known in mathematics [7] but not in engineering. It is based on the following concept: A function $\mathbf{f}(\mathbf{x})$ is said to be a *contraction* if

$$\|\mathbf{f}(\mathbf{x})\| < \|\mathbf{x}\| \qquad (\mathbf{f}(\mathbf{0}) = \mathbf{0})$$

for some set of values of $\mathbf{x} \neq \mathbf{0}$ and *some* norm. With this in mind, we can state

Theorem 4*. *Consider the discrete-time, free, stationary, dynamic system*

$$\mathbf{x}(t_{k+1}) = \mathbf{h}(\mathbf{x}(t_k)) \qquad (\mathbf{h}(\mathbf{0}) = \mathbf{0}) \qquad (5^*\text{-FS})$$

ASSUME *that* $\mathbf{h}$ *is a contraction for all* $\mathbf{x}$ *and some norm.*

THEN *the system (5*-FS) is asymptotically stable in the large, and one of its Lyapunov functions is*

$$V(\mathbf{x}) = \|\mathbf{x}\| \qquad (2)$$

---

[7] If $\mathbf{A}$, $\mathbf{B}$ are two matrices, we write $\mathbf{A} < \mathbf{B}[\mathbf{A} \leq \mathbf{B}]$ to express the fact that $\mathbf{B} - \mathbf{A}$ is positive definite [semidefinite].

[8] We write $\mathbf{x}'\mathbf{A}\mathbf{x} = \|\mathbf{x}\|^2_{\mathbf{A}}$.

[9] I.e., the infinite sum (17*) converges.

PROOF: Computing ΔV and assuming $t_{k+1} - t_k = 1$ (see footnote 4), we find that

$$\Delta V = \|\mathbf{h}(\mathbf{x})\| - \|\mathbf{x}\| \tag{3}$$

which is negative definite by definition of a contraction. Q.E.D.

Example 2. Conditions for Contraction. Consider the system

$$\mathbf{x}(t_{k+1}) = \mathbf{H}(\mathbf{x}(t_k))\mathbf{x}(t_k) \tag{4}$$

Assume that there are positive constants $c_1, \ldots, c_n$ such that either

$$(a) \quad \max_i \left\{ \sum_{j=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \right\} < 1 \quad \text{for all} \quad \mathbf{x}$$

or

$$(b) \quad \max_j \left\{ \sum_{i=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \right\} < 1 \quad \text{for all} \quad \mathbf{x}$$

Then in either case $\mathbf{H}(\mathbf{x})\mathbf{x}$ is a contraction for all $\mathbf{x}$ and therefore the system (4) is asymptotically stable in the large.

In case (a), define the norm by

$$\|\mathbf{x}\| = \max_i \{c_i|x_i|\}$$

Then

$$\begin{aligned} \|\mathbf{H}(\mathbf{x})\mathbf{x}\| &= \max_i \left\{ c_i \left| \sum_{j=1}^{n} h_{ij}(\mathbf{x}) x_j \right| \right\} \\ &\leqq \max_i \left\{ \sum_{j=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \cdot c_j |x_j| \right\} \\ &\leqq \max_i \left\{ \sum_{j=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \right\} \cdot \max_j \{c_j|x_j|\} \end{aligned}$$

which verifies that $\mathbf{H}(\mathbf{x})\mathbf{x}$ is a contraction. Note that in this norm, the contours V = constant = c_0 are cubes with edges of length $2c_0c_i^{-1}$.

In case (b), define the norm by

$$\|\mathbf{x}\| = \sum_{i=1}^{n} c_i|x_i|$$

Then

$$\begin{aligned} \|\mathbf{H}(\mathbf{x})\mathbf{x}\| &= \sum_{i=1}^{n} c_i \left| \sum_{j=1}^{n} h_{ij}(\mathbf{x}) x_j \right| \\ &\leqq \sum_{i=1}^{n} \frac{c_i}{c_j} \sum_{j=1}^{n} |h_{ij}(\mathbf{x})| \cdot c_j |x_j| \\ &\leqq \max_j \left\{ \sum_{i=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \right\} \cdot \sum_{j=1}^{n} c_j |x_j| \end{aligned}$$

which yields the desired result.

It is clear that whether or not a given function is a contraction depends on the norm chosen. This is where ingenuity is needed.

8—Estimation of Transient Behavior

By regarding V as a measure of "distance" in state space, the estimation of transient behavior proceeds in a manner which is completely analogous to that given in the companion paper.[3] Only the details of some of the calculations are different. This is illustrated by the analog of Example 15:

Example 15*. Effect of Perturbations and Parameter Variations in the Linear Stationary Case. Consider the system $(t_{k+1} - t_k = 1)$

$$\mathbf{x}(t_{k+1}) = \mathbf{H}\mathbf{x}(t_k) + \mathbf{K}(\mathbf{x}(t_k))\mathbf{x}(t_k) + \mathbf{u}(t_k) \tag{60*}$$

where

$$\left.\begin{aligned} \|\mathbf{K}(\mathbf{x}(t_k))\| &\leqq c_0 \\ \|\mathbf{u}(t_k)\| &\leqq c_1 \end{aligned}\right\} \quad \text{for all} \quad t_k$$

Assuming $\mathbf{H}$ is asymptotically stable, take any symmetric, positive-definite $\mathbf{Q}$ and define a Lyapunov function

$$V(\mathbf{x}) = \|\mathbf{x}\|_{\mathbf{P}}^2$$

by means of Corollary 3.1*. Then, evidently,

$$\begin{aligned} \Delta V(\mathbf{x}) = &-\|\mathbf{x}\|_{\mathbf{Q}}^2 + 2[\mathbf{x}'(t_k)\mathbf{K}'(\mathbf{x}(t_k)) + \mathbf{u}'(t_k)]\mathbf{P}\mathbf{H}\mathbf{x}(t_k) \\ &+ [\mathbf{x}'(t_k)\mathbf{K}'(\mathbf{x}(t_k)) + \mathbf{u}'(t_k)]\mathbf{P}[\mathbf{K}(\mathbf{x}(t_k))\mathbf{x}(t_k) + \mathbf{u}(t_k)] \end{aligned} \tag{61*}$$

Employing the Schwarz inequality, (61*) may be written in the form:

$$\begin{aligned} \Delta V(\mathbf{x}) \leqq &\left[\frac{-\|\mathbf{x}\|_{\mathbf{Q}}^2 + 2c_0\|\mathbf{x}\| \cdot \|\mathbf{P}\mathbf{H}\mathbf{x}\| + c_0^2\|\mathbf{P}\| \cdot \|\mathbf{x}\|^2}{\|\mathbf{x}\|_{\mathbf{P}}^2} \right] V(\mathbf{x}) \\ &+ 2c_1 \left[\frac{\|\mathbf{P}\mathbf{H}\mathbf{x}\| + c_0\|\mathbf{P}\| \cdot \|\mathbf{x}\|}{\|\mathbf{x}\|_{\mathbf{P}}} \right] V^{1/2}(\mathbf{x}) + c_1^2\|\mathbf{P}\| \end{aligned} \tag{5}$$

Using repeatedly the relation ($\mathbf{A}$, $\mathbf{B}$ = symmetric, positive-definite)

$$\max_{\mathbf{x}} \{\|\mathbf{x}\|_{\mathbf{A}}^2/\|\mathbf{x}\|_{\mathbf{B}}^2\} = \lambda_{\max}(\mathbf{A}\mathbf{B}^{-1})$$

each one of the terms in (5) may be maximized separately, which leads to the following:

$$\begin{aligned} \Delta V(\mathbf{x}) \leqq &[-e^{2\eta T} + 2c_0c_3c_4 + c_0^2c_2^2c_3^2]V \\ &+ 2c_1[c_4 + c_0c_2^2c_3]V^{1/2} + c_1^2c_3^2 \end{aligned} \tag{6}$$

where

$$e^{2\eta T} = \lambda_{\min}(\mathbf{Q}\mathbf{P}^{-1}); \quad T = t_{k+1} - t_k$$

η being an estimate of the time constant of the system similar to that discussed in the companion paper;[3] and

$$\begin{aligned} c_2 &= \lambda_{\max}^{1/2}(\mathbf{P}) \\ c_3 &= [\lambda_{\min}(\mathbf{P})]^{-1/2} \\ c_4 &= \lambda_{\max}^{1/2}(\mathbf{H}'\mathbf{P}^2\mathbf{H}\mathbf{P}^{-1}) \end{aligned}$$

From (6) we see that $\Delta V(\mathbf{x})$ will be negative if

$$\|\mathbf{x}\| > \frac{c_1\left[c_0c_2^2c_3 + c_4 + \sqrt{e^{2\eta T}c_3^2 + c_4^2}\right]}{e^{2\eta T} - 2c_0c_3c_4 - c_0^2c_2^2c_3^2}$$

provided that the denominator of the foregoing expression is positive, which is just the condition for asymptotic stability when $c_1 = 0$.

9—Relations With System Optimization

Similarly to the continuous-time case, define a *performance index* in the stationary case by

$$V(\mathbf{x}) = \sum_{t_k=0}^{\infty} \rho(\boldsymbol{\phi}(t_k;\ \mathbf{x}, 0)) \tag{65*}$$

where ρ is the *error criterion*.

For exactly the same reasons as before, we are led to

Theorem 5*. (Kalman [8]) *Consider a free, linear, stationary dynamic system with an equilibrium state at the origin and* ASSUME:

(i) *The error criterion $\rho(\mathbf{x})$ is positive definite and $\rho(\mathbf{0}) = 0$;*

(ii) *The performance index* $V(\mathbf{x})$ *defined by* (65) *is finite in some neighborhood of the origin.*

THEN $\mathbf{x}_s = \mathbf{0}$ *is asymptotically stable.*

The entire discussion of this theorem in the companion paper[3] is applicable here too. We add only one remark:

In case $\rho(\mathbf{x}) = x_1^2$ (x_1 being the output of the system), the performance index (65*) is obviously minimized for arbitrary $\mathbf{x}(0)$ if $x_1(t_k) \equiv 0$ for all $t_k > 0$. As Bergen and Ragazzini have shown [9], this can indeed be accomplished by suitable digital compensation of single-input systems. As they also pointed out, however, their design procedure (which is unique) does not necessarily result in a *stable* system.

The situation is as follows: If $H(z)$ is the open-loop pulse transfer function of the plant to be controlled including the zero-order hold, then the transfer function $D(z)$ of the Bergen-Ragazzini controller must contain poles to cancel the zeros of $H(z)$ because of the condition $D(z)H(z) = z - 1$. This cancellation is never exact, and if $H(z)$ contains zeros outside the region $|z| \leqq 1$, then elementary root-locus arguments show that the resulting system will have a pole outside the unit circle and hence be unstable, no matter how nearly the zero is canceled. This can happen quite readily even if the plant is open-loop stable; in such cases Bergen and Ragazzini recommend a modified procedure.

It is sometimes stated that the Bergen-Ragazzini design as just sketched fails because of the impossibility of perfect cancellation. This is misleading and erroneous, for *even with perfect cancellation the Bergen-Ragazzini design is not asymptotically stable when* $H(z)$ *contains zeros outside the unit circle.*

In fact, suppose the contrary; i.e., under perfect cancellation the system is asymptotically stable. An imperfect cancellation is readily seen to be equivalent to a perturbation of the equations of the system, as in (60*), But, by Example 15*, if c_0 is sufficiently small (i.e., good but not necessarily perfect cancellation) then the system remains asymptotically stable. *In other words, asymptotic stability is a system property which cannot be destroyed by small perturbations in the coefficients.* But this contradicts the foregoing conclusions about imperfect cancellation.

The fallacy in the conventional reasoning is that one cannot, in general, conclude that a stable response to a step input implies stability of motions starting at arbitrary initial states (see also condition (ii) of Theorem 3*).

We have here a particularly simple example of the fact that minimization of performance indexes based on semidefinite error criteria is not necessarily free of stability difficulties.

10—Application to Design

In the corresponding section of the companion paper,[3] we considered the problem of designing a regulating system for the plant governed by the equations

$$d\mathbf{x}/dt = \mathbf{F}\mathbf{x} + \mathbf{D}\mathbf{u}(t) \tag{7}$$

where the control variables are subject to the constraints

$$|u_i(t)| \leqq c_i < \infty \quad (i = 1, \ldots, m) \tag{8}$$

If the controller is to be a sampled-data system, it is convenient to transform (7) first into the discrete-time equivalent form

$$\mathbf{x}(t_{k+1}) = \mathbf{H}\mathbf{x}(t_k) + \mathbf{\Delta}\mathbf{u}(t_k) \tag{70*}$$

where it was assumed that $t_{k+1} - t_k = T =$ constant for all k, and

$$\mathbf{H} = \exp T\mathbf{F}; \quad \mathbf{\Delta} = \int_0^T (\exp \tau\mathbf{F})\mathbf{D}\, d\tau$$

Assume that $\mathbf{F}$ (and therefore $\mathbf{H}$) is asymptotically stable and given any symmetric, positive-definite $\mathbf{Q}$ find the corresponding symmetric, positive-definite matrix $\mathbf{P}$ by means of Corollary 3.1*. Then

$$V(\mathbf{x}) = \|\mathbf{x}\|_{\mathbf{P}}^2$$

is clearly a Lyapunov function for the free part of system (70*). The difference of V along a motion of (70*) is then

$$\mathbf{\Delta}V(\mathbf{x}(t_k), t_k) = T^{-1}[-\|\mathbf{x}(t_k)\|_{\mathbf{Q}}^2 + 2\mathbf{u}'(t_k)\mathbf{\Delta}'\mathbf{P}\mathbf{H}\mathbf{x}(t_k) + \mathbf{u}'(t_k)\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta}\mathbf{u}(t_k)] \tag{9}$$

An approximately optimal and practically useful design is obtained by selecting $\mathbf{u}(t_k)$ at each sampling instant t_k in such a way as to minimize ΔV.

It is clear that the right-hand side of (9) is a quadratic function in the components of $\mathbf{u}$. Disregarding at first the limits (8) on the components of $\mathbf{u}$, $\partial\Delta V/\partial\mathbf{u} = \mathbf{0}$ if $\mathbf{u}$ is given by

$$\mathbf{u}^\circ(t_k) = -(\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta})^{-1}\mathbf{\Delta}'\mathbf{P}\mathbf{H}\mathbf{x}(t_k) \tag{10}$$

The matrix $\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta}$ is obviously positive semidefinite since $\mathbf{P}$ is positive definite. If $\mathbf{\Delta}$ is 1-to-1 (i.e., $\mathbf{\Delta}\mathbf{u} = \mathbf{0}$ implies $\mathbf{u} = \mathbf{0}$ or, equivalently, the columns of $\mathbf{\Delta}$ are a linearly independent set), then $\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta}$ is positive definite so that the inverse required in formula (10) actually exists. Moreover, positive definiteness of $\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta}$ implies also that (9) has a true minimum at the point where the derivative with respect to $\mathbf{u}$ vanishes.

If we now take into account (8), it follows that $\mathbf{u}$ should be set as close to $\mathbf{u}^\circ$ as allowed by the constraints. Hence

$$\begin{aligned} u_i^*(t_k) &= -c_i \operatorname{sat}\,[c_i^{-1}(\mathbf{\Delta}'\mathbf{P}\mathbf{\Delta})^{-1}\mathbf{\Delta}'\mathbf{P}\mathbf{H}\mathbf{x}(t_k)]_i \quad (i = 1, \ldots, m) \\ \mathbf{u}^*(t) &= \mathbf{u}^*(t_k), \qquad t_k \leqq t < t_{k+1} \end{aligned} \tag{72*}$$

where

$$\operatorname{sat} x = \begin{cases} +1, & x > 1 \\ x, & |x| \leqq 1 \\ -1, & x < -1 \end{cases}$$

The difficulties encountered in the companion paper[3] in proving asymptotic stability in the large when $\mathbf{F}$ and $\mathbf{D}$ are not known exactly and when some state variables are not measurable do not arise here because the function *sat* changes by small amounts for small perturbations in the argument.

Theorem 6*. *Let* $\mathbf{x}$ *be expressed uniquely as* $\mathbf{y} + \mathbf{z}$ *where* $\mathbf{y}$ *denotes the measurable and* $\mathbf{z}$ *the unmeasurable state variables;* $\hat{\mathbf{z}}$ *is the predicted value of* $\mathbf{z}$. *Let* $\hat{\mathbf{H}}$, $\hat{\mathbf{\Delta}}$ *denote the available estimates of* $\mathbf{H}$, $\mathbf{\Delta}$ *and* $\hat{\mathbf{P}} = \mathbf{P}(\hat{\mathbf{H}}, \mathbf{Q})$. IF

(i) $|\lambda_i(\mathbf{H}_{22})| < 1$ *for all* i *and*
(ii) $\mathbf{H} - \hat{\mathbf{H}}$, $\mathbf{\Delta} - \hat{\mathbf{\Delta}}$ *are sufficiently small*

THEN *the over-all system*

$$\mathbf{x}(t_{k+1}) = \mathbf{H}\mathbf{x}(t_k) + \mathbf{\Delta}\mathbf{u}^*(\hat{\mathbf{H}}, \hat{\mathbf{\Delta}}, \hat{\mathbf{P}}, \mathbf{y}(t_k) + \hat{\mathbf{z}}(t_k))$$

$$\hat{\mathbf{z}}(t_{k+1}) = \hat{\mathbf{H}}_{21}\mathbf{y}(t_k) + \hat{\mathbf{H}}_{22}\hat{\mathbf{z}}(t_k) + \hat{\mathbf{\Delta}}_2\mathbf{u}^*(\hat{\mathbf{H}}, \hat{\mathbf{\Delta}}, \hat{\mathbf{P}}, \mathbf{y}(t_k) + \hat{\mathbf{z}}(t_k))$$

is asymptotically stable in the large. (In the foregoing, $\mathbf{u}^*$ denotes the function defined by (72*); $\mathbf{H}_{21}$, $\mathbf{H}_{22}$ are the lower two submatrices obtained by partitioning $\mathbf{H}$ into four parts, in accordance with decomposing $\mathbf{x}$ as the direct sum of $\mathbf{y}$ and $\mathbf{z}$.)

PROOF. Assume first that $\mathbf{H}$, $\mathbf{\Delta}$ are known exactly. Using the notation $\tilde{\mathbf{z}} = \mathbf{z} - \hat{\mathbf{z}}$ for the error in predicting the values of the unmeasurable state variables, the over-all system equations become:

$$\left.\begin{aligned} \mathbf{x}(t_{k+1}) &= \mathbf{H}\mathbf{x}(t_k) + \mathbf{\Delta}\mathbf{u}^*(\mathbf{x}(t_k) + \tilde{\mathbf{z}}(t)) \\ \tilde{\mathbf{z}}(t_{k+1}) &= \mathbf{H}_{22}\tilde{\mathbf{z}}(t_k) \end{aligned}\right\} \tag{11}$$

Let $\mathbf{N}$ be a symmetric, positive-definite matrix. Using assumption (i), find the corresponding symmetric, positive-definite matrix $\mathbf{M}$ via Corollary 3.1*. Define the Lyapunov function

$$V(\mathbf{x}, \tilde{\mathbf{z}}) = \|\mathbf{x}\|_{\mathbf{P}}^2 + \|\tilde{\mathbf{z}}\|_{\mathbf{M}}^2 \tag{12}$$

Using the abbreviation $\mathbf{W} = \mathbf{H}'\mathbf{P}\Delta(\Delta'\mathbf{P}\Delta)^{-1}\Delta'\mathbf{P}\mathbf{H}$, we find after simple calculations

$$\Delta V(\mathbf{x}, \tilde{\mathbf{x}}) \leq T^{-1}\{-\lambda_{\min}(\mathbf{Q})\|\mathbf{x}\|^2 - \lambda_{\min}(\mathbf{N})\|\tilde{\mathbf{x}}\|^2 + \|\mathbf{W}\|[2\|\mathbf{x}\|.\|\tilde{\mathbf{x}}\| + \|\tilde{\mathbf{x}}\|^2]\}$$

This expression is negative definite if

$$-\lambda_{\min}(\mathbf{Q})[\lambda_{\min}(\mathbf{N}) - \|\mathbf{W}\|] + \|\mathbf{W}\|^2 < 0;$$

the foregoing condition is satisfied if $\lambda_{\min}(\mathbf{N})$ is sufficiently large, which can be brought about by making $\mathbf{N}$ large.

It is now clear that (11) is asymptotically stable in the large, with a Lyapunov function of the form (12) and

$$\Delta V(\mathbf{x}, \tilde{\mathbf{x}}) \leq -k(\|\mathbf{x}\|^2 + \|\tilde{\mathbf{x}}\|^2)$$

where $0 < k < T^{-1}\lambda_{\min}(\mathbf{Q})$.

Inexact knowledge of $\mathbf{H}$ and Δ adds perturbation terms to (11), analogous to the terms $\mathbf{K}(\mathbf{x}(t_k))\mathbf{x}(t_k)$ considered in Example 15*. By assumption (ii) and the results of Example 15*, the conclusion of Theorem 6* follows. Q. E. D.

References

1 J. R. Ragazzini and G. F. Franklin, "Sampled-Data Control Systems" McGraw-Hill Book Company, New York, N. Y., 1958.

2 Ta Li, "The Stability Problem in Difference Equations" (in German), *Acta Mathematica*, vol. 63, 1934, pp. 99–141.

3 W. Hahn, "On the Application of the Method of Lyapunov to Difference Equations" (in German), *Mathematische Annalen*, vol. 136, 1958, pp. 403–441.

4 W. Hahn, "Theory and Application of the Direct Method of Lyapunov (in German), *Erg. der Math.*, vol. 22, Springer, 1959.

5 R. E. Kalman and J. E. Bertram, "A Unified Approach to the Theory of Sampling Systems," *Journal of The Franklin Institute*, vol. 267, 1959, pp. 405–436.

6 R. E. Kalman, "Nonlinear Aspects of Sampled-Data Control Systems," Proceedings, Second, Symposium Nonlinear Circuit Theory, Polytechnic Institute of Brooklyn, 1956, pp. 273–313.

7 A. N. Kolmogorov and S. V. Fomin, "Elements of the Theory of Functions and Functional Analysis," vol. 1. Translated by L. F. Boron, Graylock, Rochester, N. Y., 1957.

8 R. E. Kalman, "Dynamic Optimization of Linear Control Systems. I—Theory," to be published.

9 A. R. Bergen and J. R. Ragazzini, "Sampled-Data Processing Techniques for Feedback Control Systems," *Trans. AIEE*, vol. 73, II, 1954, pp. 236–247.

G. SACHS
Associate Director
Mem. ASME

J. G. SESSLER
Senior Project Engineer

R. F. PRAY
Research Associate

T. H. YEH
Research Associate
Syracuse University Research Institute, Syracuse, New York

Relations Between the Notch Tensile Strength of Cylindrical and Prismatic Specimens of Titanium Alloys and Heat-Treated Steels[1]

The basic relations which govern the behavior of titanium alloys in the presence of stress concentrations, as caused by notching, were investigated for cylindrical and prismatic tension test specimens. Notch tensile strengths were determined, at room temperature, for 4340 steel heat-treated to various strength levels, and for the titanium alloys 6Al-4V and 5Al-1.5Fe-1.4Cr-1.2Mo (Ti-155A). Some effects of notch geometry, specimen shape, type of loading, material history, and test temperature were included in the investigation.

The information acquired from these studies leads to recommendations pertaining to the design of notch tensile specimens for evaluating the relative notch sensitivity of titanium sheet alloys.

General Introduction

STRUCTURAL AIRCRAFT PARTS made from high bred alloys are known to behave occasionally in a very brittle manner. Many factors have been recognized as contributing to such premature failures and these relations have been extensively studied for heat-treated steel forgings at room temperature [1][2] and for various heat resisting alloys at elevated temperature [2]. In both instances, sharply notched cylindrical tension specimens, developed for this purpose [3, 4] are extensively used as a means of evaluating embrittling factors.

In contrast, no such basic information and very little practical test data are available on the tendency toward embrittlement of shapes, other than cylindrical, and particularly on that of sheet [2, 5 to 9]. The small amount of such information does not provide any definite basis for the evaluation of embrittlement. This also applies to some data on tubing [10].

The primary aim of such tests is to obtain a stress or "strength" value which indicates a considerable loss in load carrying capacity within a certain area of service variables. This deficiency is revealed by comparing the test value, usually the "notch strength" of a sharply notched specimen, with a reference value for the same material under conditions where this material would behave in a ductile manner.

The objectives of the work described in this paper were to establish such basic data for titanium alloy sheet [5 to 9]. As a first step in this direction, it was considered necessary to explore two major sets of relations. First, the relations between the behavior of titanium alloys to that of a material where the embrittlement phenomena are well recognized must be established. Such data, at room temperature, are primarily available for cylindrical specimens of heat-treated steels [3, 4, 14 to 18].

A few tests on titanium alloys in the form of cylindrical specimens [19 to 23] indicate that this group of materials behaves in a manner very similar to heat-treated steels.

Therefore the first part of this paper concerns a comparison of the effects of a number of factors on the notch strength of both steels and titanium alloys.

Second, the relations between the behavior of flat specimens to that of cylindrical specimens must be established. Because of the complete lack of this type of data, it appeared desirable to study again both titanium alloys and heat-treated steels.

All materials can be classified as either "notch insensitive" or "notch sensitive." Notch insensitive metals possess a notch strength that is generally higher than the ultimate tensile strength by an amount that depends upon the notch geometry. In the case of round specimens, the notch strength ratio of sharply notched cylindrical specimens which are of primary interest is generally increased as the "notch depth," or area removed by notching, becomes greater. For example, for 30 per cent notch depth, the ratio of notch strength to tensile strength, or "notch strength ratio" (of many "notch ductile" materials), is about 1.3, and for 50 per cent notch depth about 1.5. This applies both to heat-treated steels [3, 4] and to titanium alloys [20, 21].

These relations are unknown for shapes other than cylindrical, which are subjected to concentric loading.

A notch sensitive, or "notch brittle," material generally exhibits a notch strength ratio that is below that of notch ductile materials, if the specimen is properly designed, i.e., if it is not too small and the notch is deep and sharp.

The notch strength ratio of such a specimen, therefore, is a measure of the sensitivity of this material to suffer a loss in load carrying capacity if the part contains areas of high stress concentration. However, it must be compared to the value for a similar, notch ductile alloy condition. Also, notch sensitivity usually occurs only in certain ranges of geometrical variables and cannot be revealed if the tests are outside of this area, e.g., if the notch, of the section variations, are rather mild.

[1] The work described in this paper was performed under Syracuse University Research Institute Contract NOas 57-483-d with the Navy Department, Bureau of Aeronautics.

[2] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 26, 1959. Paper No. 59—Met-2.

Furthermore, when cylindrical specimens are tested concentrically they are considered as mildly notch sensitive if the notch strength ratio is greater than unity, but as severely notch sensitive if the notch strength ratio is less than unity. However, this is an arbitrary distinction, and it is very doubtful that it can be applied to flat (sheet) specimens.

This paper presents only experimental evidence. Fundamental evaluation of its content will be postponed until further test data become available on sheet specimens of various dimensions. However, in order to facilitate the use of the data presented to date, a graph is presented in Appendix II which shows the elastic stress concentrations for the major geometries used in this paper.

Experimental Procedure

Materials. The steel and titanium alloys included in the investigation are presented in Table 1.

Specimens. Three basic types of test specimens were used for the investigation. Cylindrical, square, and rectangular cross-section specimens are shown in Figs. 1(*a*, *b*, and *c*), respectively. All of the specimens were machined from original stock with the axis of the specimen parallel to the rolling direction of the material. The 4340 steel cylindrical and prismatic specimens were rough machined to within 0.020 in. of the final dimensions, heat-treated according to Table 1, and then finish ground to size. The titanium specimens were heat-treated in the form of blanks and then machined to the final dimensions after heat-treatment. Cylindrical specimens were turned in a lathe and flat specimens were milled to the desired shape.

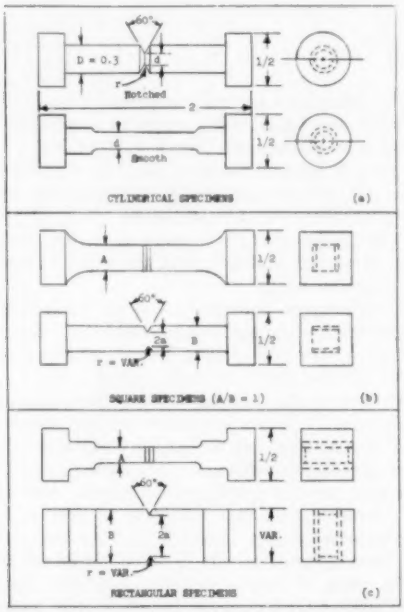


Fig. 1 Cylindrical, square, and rectangular specimens used for the investigation

Notches in cylindrical specimens were machined with a single point, carbide-tipped lathe tool which was carefully ground to the desired notch radius. The flat specimens were notched with milling cutters and carbide-tipped swing tools of selected radii. In addition, a few notches were made by grinding (4340 specimens only) to compare the behavior of ground notches to that of machined notches of like radii and depth. No significant difference in their behavior was observed. The notches in titanium specimens were all obtained by machining.

The same cylindrical smooth and notched specimens, as shown in Fig. 1(*a*), were used for both room temperature (and below) and elevated temperature tensile and stress-rupture tests on heat-treated Ti-6Al-4V and Ti-155A.

Test Procedure. The tensile tests on both steel and titanium alloys were performed in a 60,000-pound capacity Baldwin-Southwark Universal Testing Machine, at a crosshead speed of 0.005 in. per minute. A special guiding fixture, as shown in Fig. 2, was used to insure the concentricity of axial loading that is required for the notch tensile testing of brittle materials at room temperature and below. This fixture was not used for tensile tests at elevated temperatures, but was replaced by a self-aligning system with universal joints incorporated at the upper and lower attachment points. The specimens were heated in a vertical-muffle resistance furnace and the desired test temperature was controlled to within ±5 F by means of Brown Electronik Temperature Control Units and Chromel-Alumel thermocouples attached to the specimen.

Conventional equipment and techniques were used for the performance of stress-rupture tests. These procedures have been adequately described elsewhere [24] and need not be further elaborated upon here.

PART I

Test Results on Cylindrical Specimens

Introduction. The relations which govern the notch strength of cylindrical specimens have been comparatively well established for heat-treated steels. The steel used primarily in previous systematic investigations was 3140 [4, 11]. Extensive tests have been also performed on 4340 [15 to 18].

To establish more definitely the effects of several factors, tests were performed on 4340, selecting a strength level of 165,000 psi as a notch ductile and 270,000 psi as a notch brittle condition.

The effects of a number of variables on the notch strength of cylindrical specimens are also known for several titanium alloys [19 to 23]. These range, regarding their notch sensitivity, from notch ductile to severely notch sensitive conditions. Some of

Table 1 Source, heat-treatment, and tensile strengths of the investigated alloys

Alloy	Heat-treatment	Tensile strength, psi	Original shape of stock	Supplied by
SAE 4340	1550 F-1 hr-oil 500 F-1 hr-air	270,000	Hot rolled 4-in. diam bar	Rotary Electric Steel Co.
SAE 4340	1550 F-1 hr-oil 1000 F-4 hr-air	165,000	Hot rolled 4-in. diam bar	Rotary Electric Steel Co.
Ti-6Al-4V (DOD alloy)	1725 F-10 min-water 900 F-8 hr-air	165,000	$^1/_2$-in. plate	Titanium Metals Corp.
Ti-5Al-1.5Fe 1.4Cr-1.2Mo (Ti-155A)	1700 F-1 hr-water 900 F-8 hr-air	220,000	$^3/_4$-in. diam bar	Titanium Metals Corp.

NOTE: 1 The titanium alloys were solution treated in an argon atmosphere.
2 Water quench time was less than 5 seconds.

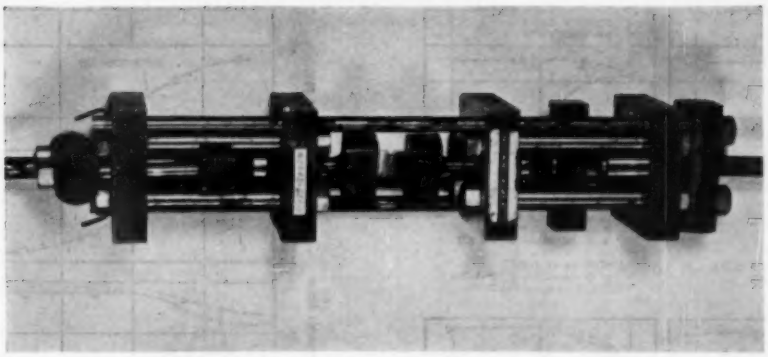


Fig. 2 Photograph of concentric guiding fixture used for the performance of tensile tests at room temperature

these data relate to Ti-6Al-4V, heat-treated to various strength levels, which appears only slightly notch sensitive, and to Ti-5Al-1.5Fe-1.4Cr-1.2Mo (Ti-155A), which is found to be highly notch brittle, if heated to strength levels above 200,000 psi [22].

This latter alloy, heat-treated to a strength level of 220,000 psi, was selected as representative of a notch brittle titanium alloy (at room temperature) to establish such relations as the effects of notch depth and notch sharpness.

Further tests were performed at testing temperatures, ranging from −100 to 800 F, on both Ti-6Al-4V and Ti-5Al-1.5Fe-1.4Cr-1.2Mo.

At relatively low temperatures, both steel and titanium alloys are known to become rather notch brittle. However, only titanium alloys were studied here, although the notch sensitivity of heat-treated steels, at temperatures below room temperature, is not clearly established.

At elevated temperatures, apparently all steels become notch sensitive within certain areas of strength, applied stress, temperature, and time. As a first approach to this problem for titanium alloys, short time tensile tests were made on both of the alloys mentioned and stress rupture tests were performed on Ti-6Al-4V.

Effects of Notch Depth and Notch Radius. The dependence of the notch strength ratio upon notch depth and notch radius is illustrated for the two steel conditions in Fig. 3 and for a number of titanium alloys (at room temperature) in Fig. 4. (The tests on specimens with very mild notches are neglected in these graphs.)

The notch ductile conditions of both groups of materials exhibit a nearly linear increase in notch strength with increasing notch depth. The trendline is also practically independent of the notch radius, and it is the same for steels and for titanium alloys. The notch strength ratio is close to 1.3 for 30 per cent and close to 1.5 for 50 per cent notch depth.

Notch sensitivity appears in Figs. 3 and 4 as a deviation to lower values from the trendline for notch ductile conditions. This deviation depends upon the notch depth in a rather complex manner, and possibly differently for steels and titanium alloys, respectively. The deviation also greatly increases as the notch radius r decreases. The notch sensitivity was not apparent if notch radii above about 0.025 in. were used in this study. Rather sharp notches must be used to develop notch sensitivity in these materials.

This effect is further illustrated by Fig. 5, in which the notch strength ratio is plotted as a function of notch sharpness a/r. Close agreement of steels and titanium alloys, in this respect, is again indicated by this graph.

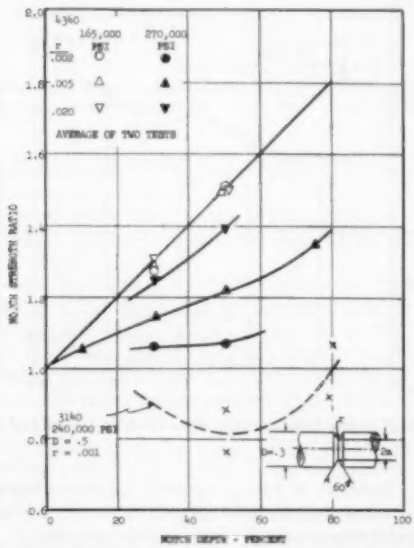


Fig. 3 Relation between notch strength ratio, notch depth, and notch radius for cylindrical specimens of 4340 steel at two strength levels

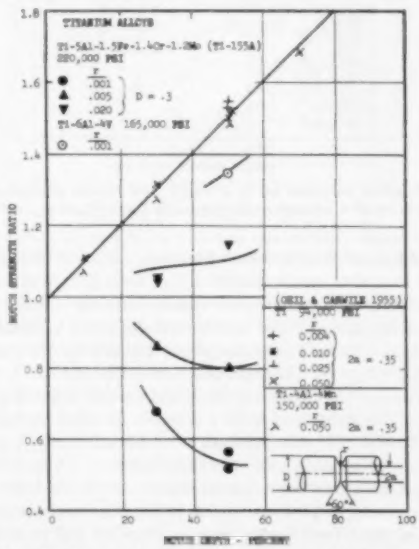


Fig. 4 Relation between notch strength ratio, notch depth, and notch radius for cylindrical specimens of several titanium alloys

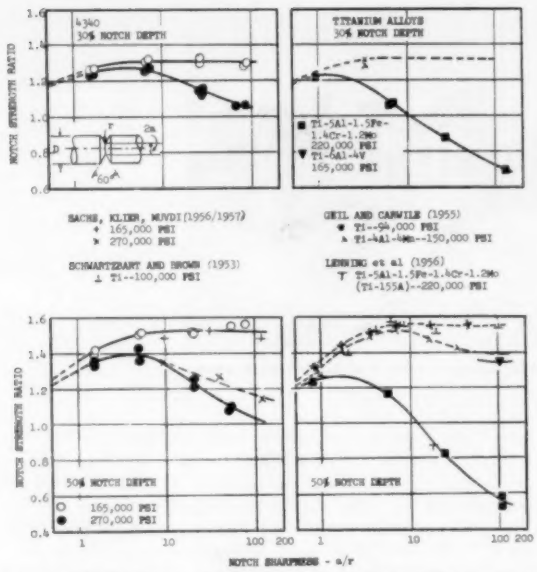


Fig. 5 Relation between notch strength ratio and notch sharpness for cylindrical specimens of 4340 steel and titanium alloys at 30 and 50 per cent notch depth

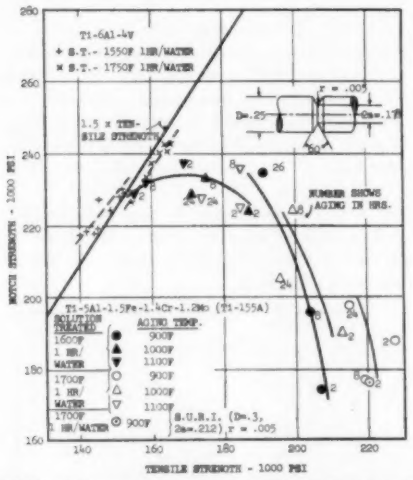


Fig. 6 Relation between notch strength and tensile strength for two titanium alloys after various heat-treatments (Lenning, et al., 1956)

Effect of Strength Level and Heat-Treatment. Figs. 3 to 5 also confirm the fact that notch sensitivity of both groups of materials can be disclosed by concentric notch tests on small (0.3 in.) cylindrical specimens if the tensile strength exceeds certain minimum values. These values are about 180,000 psi for steels and about 150,000 psi for titanium alloys, see Fig. 6.

In addition, it is known for steels that notch sensitivity varies both with chemistry and with a number of other metallurgical factors [1, 12 to 14], the strength level being identical. Similar effects can be expected for titanium alloys. Thus it appears from Fig. 6 [22] that the notch sensitivity of Ti-5Al-1.5Fe-1.4Cr-1.2Mo depends upon both solution and aging temperatures.

Results of Short Time Tensile Tests on Titanium Alloys at Different Temperatures. Steels become generally brittle at low temperatures. The extent of embrittlement revealed by notch tensile tests [15], however, is largely unknown for heat-treated steels. This problem has not been investigated here, as this would require a considerable amount of additional experimentation.

Titanium alloys are also known to become brittle in the presence of stress concentrations as the testing temperature decreases [21,

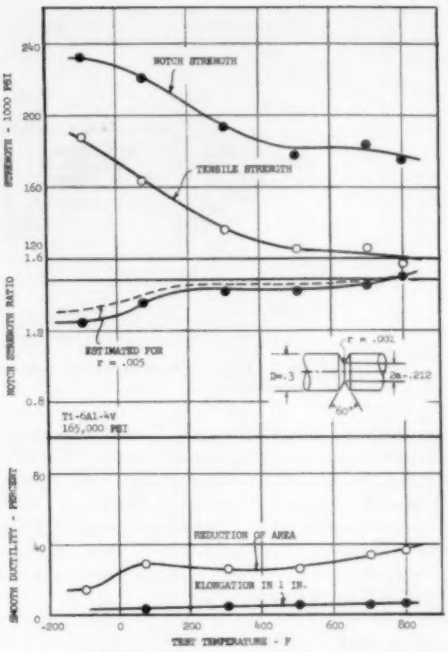


Fig. 7 Notch tensile characteristics of Ti-6Al-4V at several test temperatures

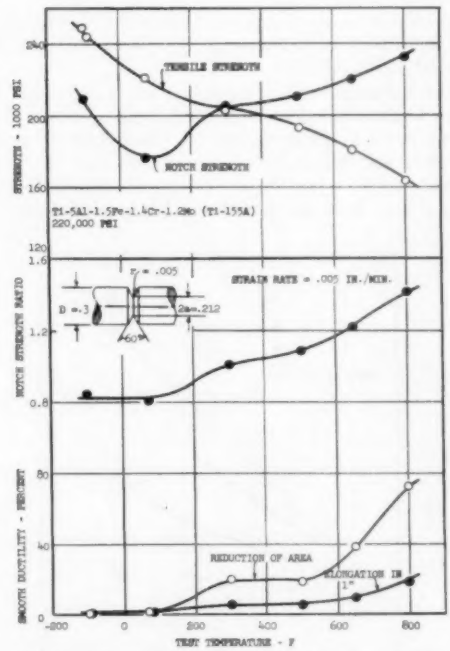


Fig. 8 Notch tensile characteristics of Ti-155A at several test temperatures

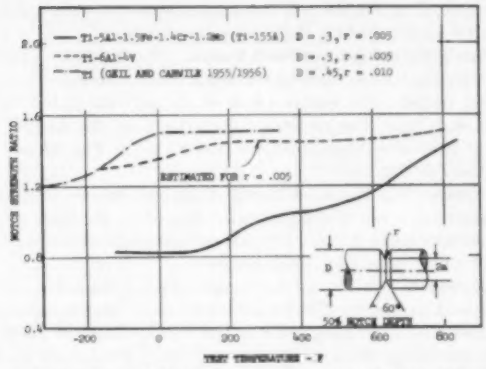

Ti-5Al-1.5Fe-1.4Cr-1.2Mo (Ti-155A) D = .3, r = .005
Ti-6Al-4V D = .3, r = .005
Ti (GEIL AND CARVILL 1955/1956) D = .45, r = .010
NOTCH STRENGTH RATIO
ESTIMATED FOR r = .005
TEST TEMPERATURE - F

Fig. 9 Comparison of the notch strength ratios of three different titanium alloys for short time tensile tests at various temperatures. (The curve for Ti-6Al-4V was obtained by interpolation.)

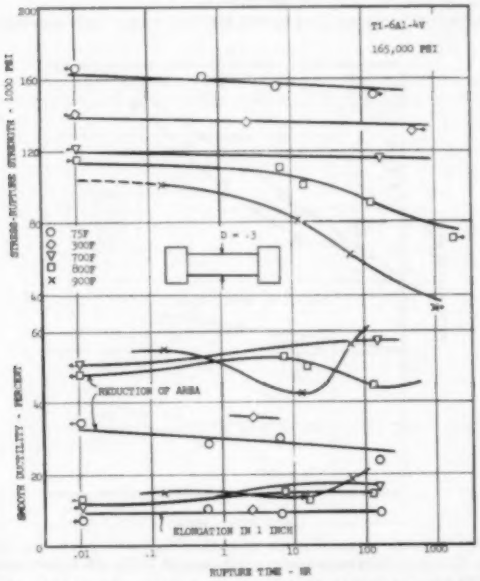


Fig. 10 Results of stress rupture tests on smooth cylindrical specimens of heat-treated Ti-6Al-4V. (All specimens were machined from ½-in. plate.)

23]. This is illustrated by Fig. 7 for Ti-6Al-4V and by Fig. 8 for Ti-5Al-1.5Fe-1.4Cr-1.2Mo.

In Fig. 9 the notch strength ratio of these alloys and of a previously studied commercially pure titanium (75,000 psi) are assembled.

Commercially pure titanium (75,000 psi) appears, according to Fig. 9 [20], to retain nearly all of its high notch ductility down to a testing temperature of −320 F. The alloy Ti-6Al-4V, heat-treated to 165,000 psi, also becomes only slightly notch sensitive, as the testing temperature decreases from about 200 to −100 F. In contrast, the notch sensitivity of Ti-5Al-1.5Fe-1.4Cr-1.2Mo (220,000 psi) continuously increases as the testing temperature progressively decreases below 800 F. However, its pronounced notch embrittlement at room temperature only slightly changes at lower temperatures.

The ductility of the two heat-treated alloys is presented in Figs. 7 and 8. Pronounced notch sensitivity, i.e., a notch strength ratio below unity, appears to be associated with a ductility of less than about 10 per cent, as measured by reduction of area.

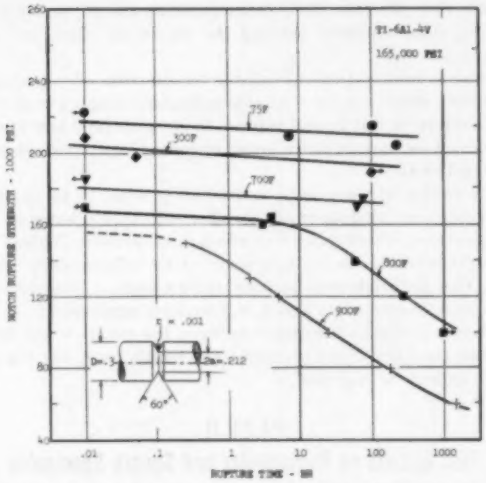


Fig. 11 Results of stress rupture tests on notched cylindrical specimens of heat-treated Ti-6Al-4V. (All specimens were machined from ½-in. plate.)

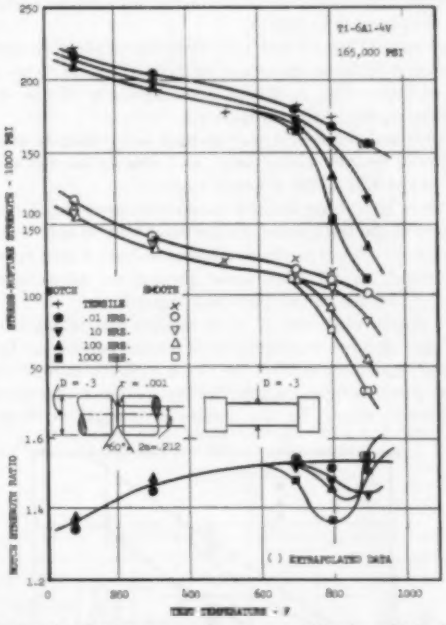


Fig. 12 Results of stress rupture tests on smooth and notched cylindrical specimens of heat-treated Ti-6Al-4V at several test temperatures

The curves in Figs. 7, 8, and 9, for the two heat-treated alloys, also indicate a certain arrest in the rate of increase with temperature of both notch strength ratio and ductility between 300 and 800 F. This also appears associated with a slight arrest in the decrease of tensile strength and notch strength with increasing temperature.

Results of Stress Rupture Tests on Ti-6Al-4V. Many heat-treated steels are known to undergo embrittlement at elevated temperatures. This phenomenon has been extensively investigated by means of stress rupture tests on both smooth and notched specimens [24 to 30]. These studies have shown that the notch strength ratio of high strength steels may be extremely low at certain combinations of temperature, applied stress, and rupture

time. This becomes more pronounced as the rupture life increases, simultaneously shifting the maximum effect to lower temperatures.

No such information is available for titanium alloys, to date. Therefore, stress rupture tests at temperatures ranging from room temperature to 900 F, and rupture lives up to 1000 hours, were performed on smooth and notched cylindrical specimens of heat-treated Ti-6Al-4V.

The results of these tests, assembled in Figs. 10 to 12, show little notch embrittlement of this alloy within 1000 hours at any temperature. Possibly, such an effect exists between 700 and 900 F, but it is within the limits of error (about ±5 per cent). However, this embrittlement appears more clearly in the ductility (reduction of area) data and it is, therefore, significant.

There is a slight discrepancy between the results of the stress rupture tests and those of short time tensile tests, see Fig. 12, which cannot be explained.

PART II

Test Results on Rectangular and Square Specimens

Introduction. The relations which govern the notch sensitivity of specimens other than cylindrical are unknown. In order to establish these basic laws, extensive studies were made on rectangular and square specimens of the two 4340 steel conditions and of heat-treated Ti-6Al-4V.

These section shapes add two variables to those encountered in tests on cylindrical specimens, namely, the ratio of the lateral specimen dimensions A/B and the magnitude of one of these dimensions, defining the "section size."

For practical reasons, it was decided to investigate shapes of equal initial cross-sectional area, and also to use the same sequence of notch radii for different shapes.

The tests have been limited to room temperature, to date.

In most of the tests, notches were machined on only two of the four surfaces. Some preliminary tests indicated that specimens with notches on all four faces yielded no advantage over those with notches on two faces, see Appendix I.

The primary objective of these studies was to establish the dependence of notch sensitivity upon the section shape. In order to measure notch sensitivity, the notch strength ratio of the material in question must be compared with that of a similar, but notch ductile, alloy. In the present investigations, 4340 steel, heat-treated to 165,000 psi, has been selected as such a ductile material for establishing the required base data.

Effects of Notch Depth and Notch Radius. The dependence of the notch strength ratio upon notch depth was investigated for two shapes, namely, the square ($A/B = 1$) and "simulated" sheet ($A/B = 1/16$). The results of these tests on the ductile and brittle 4340 steel conditions are assembled in Fig. 13 and on Ti-6Al-4V in Fig. 14.

It appears that the notch ductile 4340, heat-treated to 165,000 psi, possesses a notch strength ratio that again increases nearly linearly with notch depth. The rate of increase, however, depends greatly upon the section shape, as discussed further.

The deviation of the notch strength ratio of the notch brittle 4340, heat-treated to 270,000 psi, from these ductile base lines increases with decreasing notch radius. The maximum effect, or notch sensitivity, is found to be considerably greater for the flat shape than for the square section. This is explained by the fact that for the sharpest notch which could be obtained, $r = 0.002$ in., the notch sharpness, a/r (and stress concentration), is much larger for the flat than for the square.

Ti-6Al-4V, heat-treated to 165,000 psi, ranges between the two

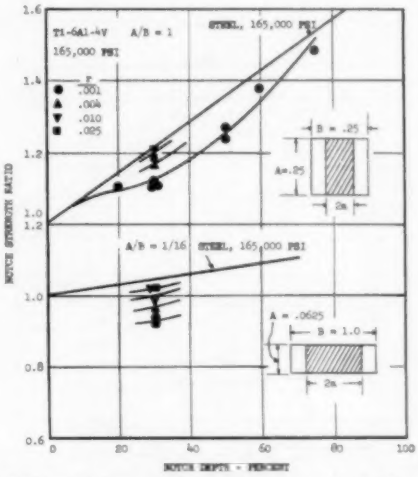


Fig. 14 Relation between notch strength ratio, notch depth, and notch radius for square and rectangular specimens of Ti-6Al-4V heat-treated to 165,000 psi

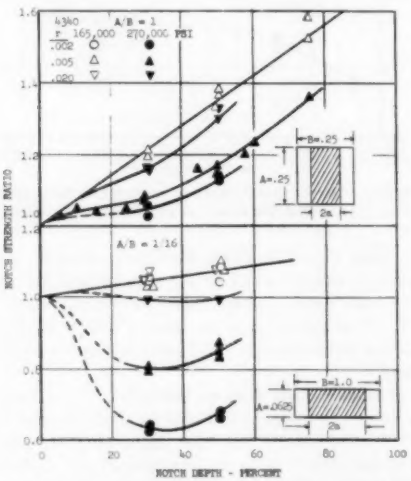


Fig. 13 Relation between notch strength ratio, notch depth, and notch radius for square and rectangular specimens of 4340 steel at two strength levels

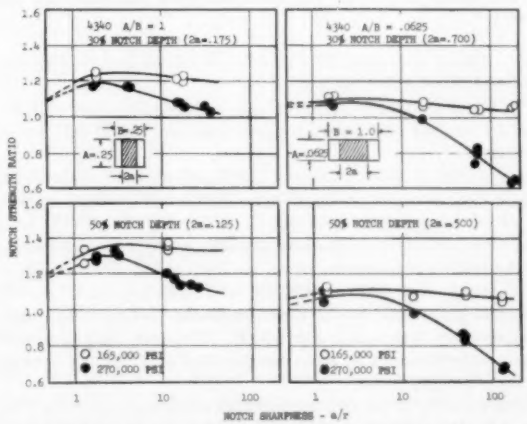


Fig. 15 Relation between notch strength ratio and notch sharpness for square and rectangular specimens of 4340 steel at 30 and 50 per cent notch depth

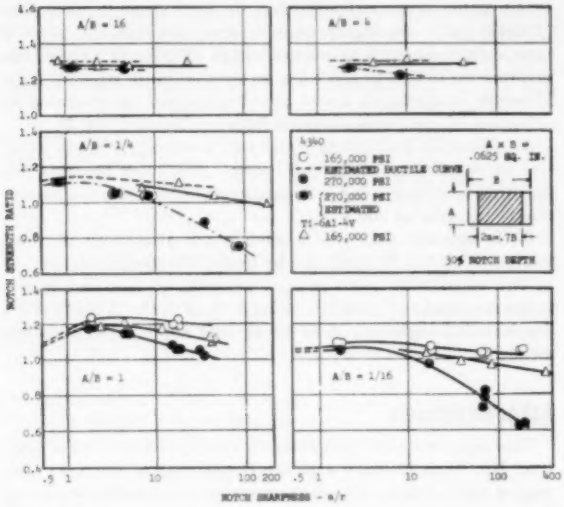


Fig. 16 Notch strength ratio as a function of notch sharpness for square and rectangular specimens of Ti-6Al-4V and 4340 steel for 30 per cent notch depth

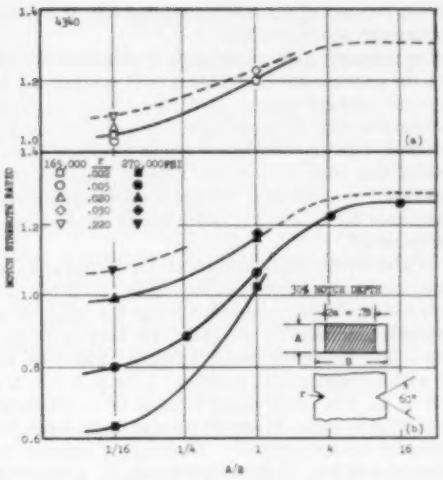


Fig. 17 Relation between notch strength ratio, A/B ratio, and notch radius for 4340 steel at two strength levels

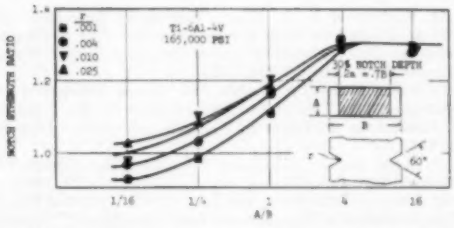


Fig. 18 Relation between notch strength ratio, A/B ratio, and notch radius for Ti-6Al-4V

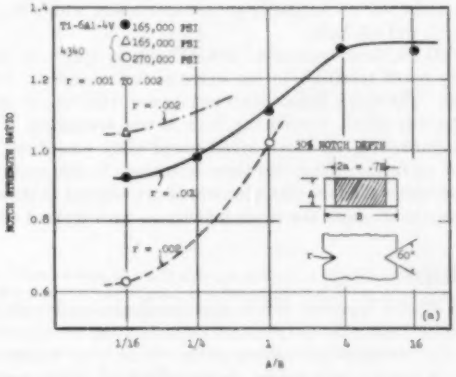


Fig. 19(a) Comparison of notch strength ratio and A/B ratio for 4340 steel and Ti-6Al-4V for 0.001 to 0.002-in. notch radius

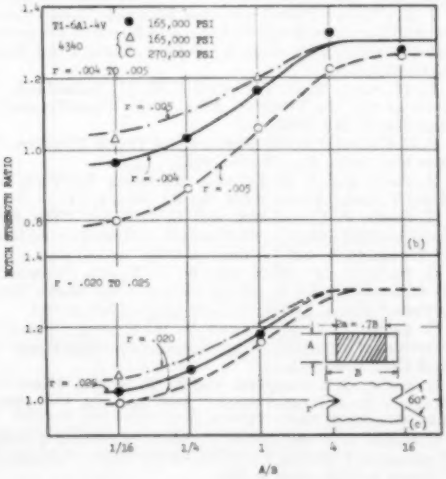


Fig. 19(b) Comparison of notch strength ratio and A/B ratio for 4340 steel and Ti-6Al-4V for 0.004 to 0.005-in. notch radius

Fig. 19(c) Comparison of notch strength ratio and A/B ratio for 4340 steel and Ti-6Al-4V for 0.020 to 0.025-in. notch radius

steel conditions. For this alloy also, the apparent notch sensitivity is slightly greater for the flat than for the square section, but this effect is less pronounced than was observed for the brittle steel condition.

However, if the data are replotted, according to Figs. 15 and 16, with notch sharpness as abscissas, the two shapes appear to be nearly equivalent in regard to notch sensitivity, at a notch sharpness of about 30. Flat sections are more sensitive to sharper notches while square sections appear slightly superior for milder notches. On the other hand, a relatively sharp notch is applied more readily to a wide strip than to a small square. This is fortunate as it renders notch tests on sheet a rather simple procedure.

Effects of Section Shape. In order to further establish the effect of geometry on notch sensitivity, five different section shapes were investigated; $A/B = 1/16$, $1/4$, 1 (square), 4, and 16.

The results of these tests on the three sections not yet discussed are also illustrated in Fig. 16, with the notch sharpness as abscissas. It is clear from this graph that the flat sections which are notched on the face, $A/B = 4$ and 16, fail to disclose notch

sensitivity, because of the mechanical inability to provide them with sufficiently sharp notches.

Further comparison of the test data is presented in Figs. 17 to 20, with the section ratio (A/B) being used as abscissas.

The notch strength ratio of 4340, heat-treated to 165,000 psi, outlines the base line (or upper limit of values) obtainable for this type of steel, Fig. 17. The notch radius has little effect on the position of this base line, except when it becomes very large (greater than $r = 0.050$ in.). Effects of such mild notches are of no significance for evaluating notch sensitivity and will not be further discussed.

Fig. 17 also reveals that the notch strengthening of a very thin, wide specimen that is notched at the edge is nearly nil. For $A/B = 1/16$, a sharp 30 per cent notch raises the notch strength of notch ductile conditions by less than 5 per cent above the tensile strength. The maximum notch strength ratio then increases rapidly with increasing A/B, to about 1.1 for $A/B = 1/4$, and 1.2 for $A/B = 1$. The notch strength ratios for notch ductile conditions at $A/B = 4$ and 16, and 30 per cent notch depth, have not been directly determined here. However, it appears from the data presented in Fig. 17 that these values are probably close to 1.3, or the same as that for cylindrical specimens.

The high notch sensitivity of 4340, heat-treated to 270,000 psi, can be disclosed, according to Fig. 17, only by testing flat specimens, notched at the edges, and nearly square specimens. Also, the notch radius must not exceed 0.010 in. for very thin specimens, and 0.005 in. for nearly square specimens, with a width between $^1/_4$ and one inch.

Ti-6Al-4V, heat-treated to 165,000 psi, is found to be only slightly notch sensitive, in the shape of flat specimens, Figs. 18 and 19. The notch radius must not exceed 0.005 in., in order to disclose this effect, beyond the limit of test scattering.

Comparison of all three investigated alloy conditions, Figs. 19(a, b, c), further illustrates these relations. It also confirms for flat specimens the fact that titanium alloys respond to stress concentrations in much the same manner as heat-treated steels.

Conclusions

The studies reported herein were conducted primarily to establish the basic relations required for evaluating the relative sensitivity of titanium sheet alloys to the effects of stress concentrations, as caused by notches. Some effects of notch geometry, specimen shape, type of loading, material history, and test temperature were investigated for both heat-treated steels and titanium alloys. These studies have resulted in the following conclusions:

1 Cylindrical and prismatic specimens of notch ductile steels and ductile titanium alloys exhibit a nearly linear increase in notch strength ratio with increasing notch depth. This rate of increase, however, depends greatly upon cross section shape. Also, these ductile trend lines are practically independent of notch radius.

2 For materials other than ductile, the deviation of notch strength ratio from the ductile trend lines to lower values increases with decreasing notch radius. The maximum effect (notch sensitivity) is nearly the same for cylindrical and flat (sheet) specimens (notched along two edges), if a sufficiently sharp notch of 30 to 50 per cent notch depth is employed for the tests. This effect is similar for both steels and titanium alloys.

3 Notch embrittlement is observed for both steels and titanium alloys if their tensile strength levels exceed about 180,000 psi and 150,000 psi, respectively. The degree of embrittlement observed depends greatly upon section shape, notch geometry, and tensile strength.

4 At room temperature, the alloy Ti-6Al-4V (165,000 psi) reveals only mild notch embrittlement, as compared to the rather severe notch sensitivity exhibited by high strength 4340 steel (270,000 psi). At higher temperatures, the titanium alloy is practically insensitive to notches within 1000 hours, at any temperature up to 900 F, according to stress-rupture strength data. However, a significant degree of embrittlement, as measured by reduction of area, was observed between 700 and 900 F.

5 The alloy Ti-155A (220,000 psi) is extremely notch sensitive at room temperature and below. This alloy becomes less sensitive with increasing test temperature and exhibits nearly ductile behavior at 800 F.

6 It appears that the relative notch sensitivity of sheet materials can best be evaluated by testing flat specimens, notched at two edges with notch radii of 0.010 in. or less, and with a maximum thickness to width ratio (A/B) of 0.25. A notch of 60-deg included angle and 30 to 50 per cent notch depth is recommended.

Acknowledgments

This paper resulted from research work performed under Contract No. NOas 57-483-d with the Department of the Navy, Bureau of Aeronautics. The authors are indebted to the Bureau of Aeronautics for granting permission to publish the results of this research. Considerable assistance in the execution of the experimentation was provided by C. Chave and R. Roeckelein.

References

1 G. Sachs and E. P. Klier, "Survey of Low-Alloy Aircraft Steels Heat Treated to High Strength Levels, Part 5—Mechanical Properties in the Presence of Stress Concentrations," Wright Air Dev. Center Tech. Rep. 53-254, 1954.

2 G. Sachs and W. F. Brown, Jr., "Notch Sensitivity of High Temperature Alloys," Syracuse Univ., Report for the Joint ASTM-ASME Committee on Effect of Temperature on the Properties of Metals, 1957.

3 G. Sachs and J. D. Lubahn, "Notched Bar Tensile Tests on Heat-Treated Low-Alloy Steels," *Trans. ASM*, vol. 32, 1943, p. 125.

4 G. Sachs, J. D. Lubahn, and L. J. Ebert, "Notched Bar Tensile Test Characteristics of Heat-Treated Low-Alloy Steels," *Trans. ASM*, vol. 33, 1944, p. 340.

5 D. S. Adams, W. M. Cottrell, and W. W. Vyvyan, "Investigation of Manufacturing Techniques for Titanium Alloy Sheet Material," Contr. No. AF 33 (600)-31606, 7th Prog. Rep., 1956.

6 H. C. Turner and C. W. Alesch, "Heat-Treatment of 6Al-4V Titanium Alloy—Notch Sensitivity," Convair, Rep. No. 9365, 1956.

7 J. K. Childs, "Determination of Materials Design Criteria for a Titanium Alloy (Ti-6Al-4V) at Room and Elevated Temperatures," Southwest Research Inst., Rep. No. P-530-7, 8, and 9 (for Wright Air Dev. Center), 1957.

8 H. H. Stier, H. C. Turner, and W. M. Sutherland, "Notch Sensitivity of 8% Mn Titanium Alloy Sheet—Heat-Treated," Convair, Rep. No. 57-243, 1957.

9 F. C. Holden, "Notch Sensitivity of Ti and Titanium Alloys," Titanium Met. Lab., Rep. No. 69, 1957.

10 G. Sachs and J. D. Lubahn, "Bursting Tests on Notched Alloy Steel Tubing," *Trans. ASM*, vol. 32, 1943, p. 71.

11 G. Sachs, J. D. Lubahn, and L. J. Ebert, "The Effects of Notches of Varying Depth on the Strength of Heat-Treated Low-Alloy Steels," *Trans. ASM*, vol. 34, 1945, p. 517.

12 G. Sachs, L. J. Ebert, and W. F. Brown, "Comparison of Various Structural Alloy Steels by Means of the Static Notch-Bar Tensile Tests," *Trans. AIME*, Iron Steel Div., 1947, p. 605.

13 G. Sachs, L. J. Ebert, and W. F. Brown, "Notch-Tensile Characteristics of a Partially Austempered, Low Alloy Steel," AIME, Iron Steel Div., 1948, p. 424.

14 G. Sachs, G. S. Sangdahl, and W. F. Brown, "New Notes on High-Strength Heat-Treated Steels," *Iron Age*, vol. 166, 1950, Nov. 23, p. 59; Nov. 30, p. 76.

15 G. Sachs, V. Weiss, and E. P. Klier, "Effects of a Number of Heat-Treating and Testing Variables on the Notch Strength of 4340 Steel," *Trans. ASTM*, 1956, p. 555.

16 J. D. Lubahn, "Effect of Section Size on Fracturing," Proceedings, 1955 Sagamore Research Conference, 1956, p. 143.

17 E. P. Klier, B. B. Muvdi, and G. Sachs, "The Static Properties of Several High-Strength Steels," *Proceedings ASTM*, vol. 57, 1957, see *Product Eng.*, vol. 28, 1957, no. 15, p. B24.

18 B. B. Muvdi, G. Sachs, and E. P. Klier, "Design Properties of High Strength Steels in the Presence of Stress Concentrations," Wright Air Dev. Center, Tech. Rep. 56-395, 1957.

19 H. Schwartzbart and W. F. Brown, "Notch-Bar Tensile Properties of Various Materials and Their Relation to the Unnotch Flow Curve and Notch Sharpness," *Trans. ASM*, vol. 46, 1954, p. 998.

20 G. W. Geil and N. L. Carwile, "Effect of Low Temperature on the Mechanical Properties of a Commercially Pure Titanium," *Journal of Research*, Nat. Bu. Stand., vol. 54, 1955, no. 2, p. 91.

21 G. W. Geil and N. L. Carwile, "Research on Effects of Prestraining and Notch Sharpness on the Notch Strength of Materials," Contr. No. AF 33(616) 55-9, 1955/1956.

22 G. A. Lenning, M. L. Greenlee, W. M. Parris, and H. R. Keasler, "The Determination of the Effect of Heat Treatment on the Elevated Temperature Stress-Stability of Titanium Alloys," Contr., Wright Air Dev. Center AF 33(616)-3638, 1956/1957.

23 E. P. Klier and N. J. Feola, "The Effects of Carbon and Nitrogen Contamination on the Notch Tensile Properties of Titanium," Symposium on Titanium, ASTM, Spec. Tech. Pub. No. 204, 1957, p. 113.

24 W. F. Brown, M. H. Jones, and D. P. Newman, "Influence of Sharp Notches on the Stress-Rupture Characteristics of Several Heat-Resisting Alloys," Symposium on Strength and Ductility of Metals at Elevated Temperatures, ASTM Spec. Tech. Pub. No. 128, 1953, p. 25.

25 E. A. Davis and M. J. Manjoine, "Effect of Notch Geometry on Rupture Strength at Elevated Temperatures," ASTM Spec. Tech. Pub. No. 128, 1952.

26 W. F. Brown, M. H. Jones, and D. P. Newman, "Influence of Sharp Notches on the Stress-Rupture Characteristics of Heat-Resisting Alloys, Part II," *Proceedings, ASTM*, vol. 53, 1953, p. 661.

27 D. P. Newman, M. H. Jones, and W. F. Brown, "Time-Temperature Dependence of the Notch Effect and Influence of Notch Depth in Stress-Rupture Tests on a Cr-Mo-V Steel," *Proceedings ASTM*, vol. 53, 1953, p. 677.

28 G. Sachs, W. F. Brown, and D. P. Newman, "The Effect of Stress Concentration on the Rupture Strength of Materials," *Zeit. Metallkunde*, vol. 44, 1953, no. 6, p. 233.

29 M. H. Jones, D. P. Newman, G. Sachs, and W. F. Brown, "Effects of Variation in Normalizing and Tempering Procedure on Stress-Rupture Strength, Creep Embrittlement and Notch Sensitivity for a Cr-Mo-V and a 17Cr-4Ni-Cu Steel," *Trans. ASM*, vol. 47, 1955, p. 926.

30 J. G. Sessler and W. F. Brown, "Notch and Smooth Bar Stress-Rupture Characteristics of Several Heat Resistant Alloys in the Temperature Range Between 600 and 1000 F," *Proceedings, ASTM*, vol. 56, 1956.

31 R. E. Peterson, "Stress Concentration Design Factors," John Wiley & Sons, New York, N. Y., 1953.

APPENDIX I

The Notch Strength of Specimens Notched on Two Sides and on Four Sides

A few tests were performed on 4340 steel, heat-treated to 270,000 psi, in which square and rectangular specimens were either notched at two sides or on all four surfaces. Various notch depths, from 10 to 80 per cent, and a constant notch radius of 0.005 in. were employed for this investigation.

The effects of notch depth were studied for square specimens with the results as shown in Fig. 20. The notch strength ratio of both types of notched square specimens (notched on 2 sides or on 4 sides) were very nearly equal at any notch depth up to 80 per cent. In addition, Fig. 20 reveals little difference in notch strength ratio between square and cylindrical specimens.

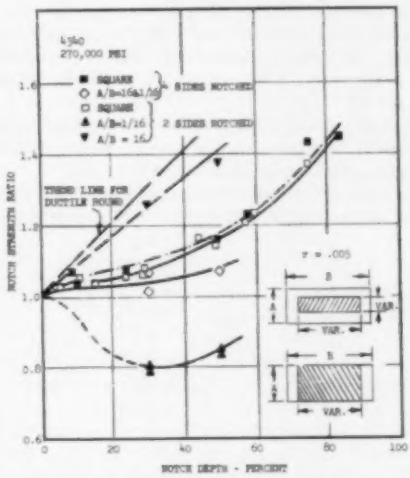


Fig. 21 Comparison of notch strength ratio as a function of notch depth for square and thin rectangular specimens of 4340 steel, having notches on either two faces or on all four faces

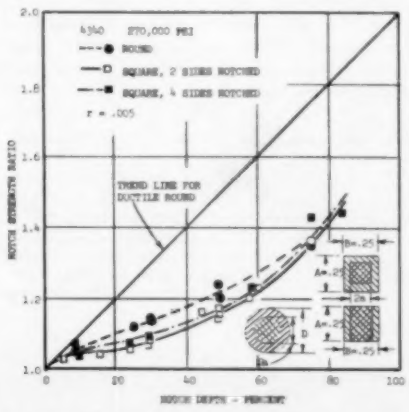


Fig. 20 Relation between notch strength ratio and notch depth for cylindrical and square specimens (notched two sides and four sides) of 4340 steel for a constant notch radius of 0.005 in.

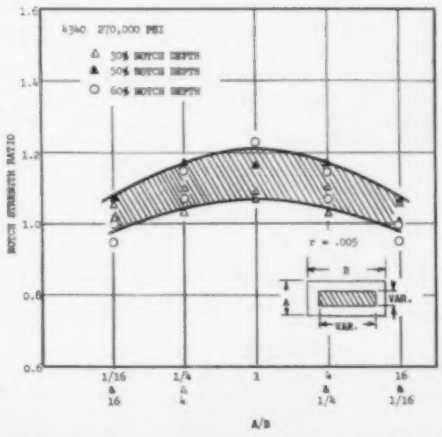


Fig. 22 Effect of section shape on the notch strength ratio of rectangular 4340 steel specimens with notches on all four faces and a constant notch radius of 0.005 in.

The small effect of section shape for specimens notched on all four faces is probably explained by the fact that their stress state is always very similar to that of cylindrical specimens. It may be assumed, therefore, that the notch strength ratio of notch ductile materials, in these shapes, is also nearly the same as that of a cylindrical specimen.

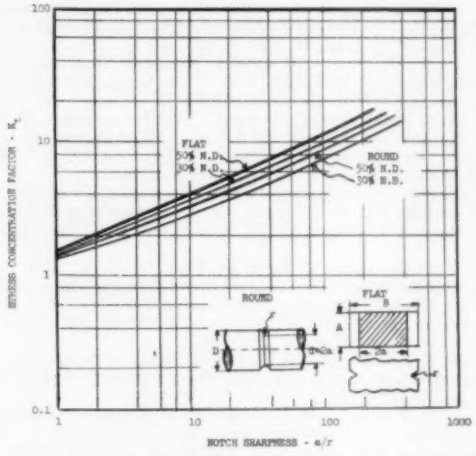
A comparison of the behavior of square and rectangular specimens (notched on 2 sides or on 4 sides) of 4340 steel (270,000 psi) is presented in Fig. 21. For thin sections, the notch strength ratio of specimens notched at two edges, as compared to those notched along two faces, is found to be greatly different. It is clear from this graph that the edge notches reveal notch sensitivity much better than notches along two faces. As discussed previously, thin sections which are notched along two faces fail to disclose notch sensitivity because of the mechanical inability to provide them with sufficiently sharp notches.

The notch strength ratios of thin specimens, notched on all four sides, are approximately halfway between the values for specimens which were notched only on two sides. Also, these values are very nearly equal to those observed for the square shape, see Figs. 21 and 22.

It is clear that the edge notch is the decisive shape factor, since, for this type of notch the notch sharpness (for a constant radius) becomes greater as the specimen width increases.

Further tests on specimens with notches on four sides were discontinued, as the considerably simpler flat specimens, notched at two edges, appear suitable for the evaluation of the notch sensitivity of sheet.

APPENDIX II



Relation between notch sharpness (a/r) and stress concentration (K_t) [31]

Nonmetallic Inclusions and Fracture Behavior of Steels

S. YUKAWA

Materials and Processes Laboratory,
Large Steam Turbine-Generator Department,
General Electric Company,
Schenectady, N. Y.

Some observations showing the adverse effects of certain kinds of nonmetallic inclusions on the strength and fracture behavior of alloy steel test specimens are presented. The observations pertain to the effects in tests at low and high temperatures and under steady and cyclic loading.

Introduction

THE PRESENCE OF A CERTAIN NUMBER of nonmetallic inclusions in structural steels as normally produced is almost inevitable. In most instances, their effect on the mechanical properties is quite innocuous. However, it is possible that under unusual or severe testing conditions, an adverse effect of inclusions on test-specimen behavior may be found.

In this paper, some observations of specific instances where the strength and fracture behavior of alloy steels were modified by the presence of inclusions are presented. The materials concerned are from large forgings of heat-treated alloy steels. The observations discussed relate to several different kinds of fracture phenomena; namely, low temperature, high temperature, and cyclic loading. The results presented are not intended to be a complete exposition of the effects of inclusions; rather, they are presented mainly to illustrate some ways in which inclusions may influence fracture behavior within the limitations of the test conditions employed.

Materials and Testing

The materials tested were taken from several large forgings of two general types. The low-temperature fracture and cyclic-loading tests were conducted on the Ni-Mo-V type of steels. For the specific Ni-Mo-V steels used, the major alloy contents were in the range 0.25 to 0.30 per cent C, 0.6 to 0.9 per cent Mn, 2.5 to 2.7 per cent Ni, 0.4 to 0.5 per cent Mo, and 0.1 per cent V. The high-temperature rupture and fracture tests were conducted on a Cr-Mo-V type of steel. The nominal amounts of the major alloys in this type of steel are 0.3 per cent C, 1 per cent Cr, 1 per cent Mo, and 0.25 per cent V. The amounts of nonmetallic elements present in the steels studied were in the normal range of values occurring in these types of steels. The forgings were heat-treated according to conventional large forging practice resulting in a tempered bainitic type of microstructure.

Since a number of different kinds of test specimens are involved in these observations, their description will be given in connection with each group of results. At this point, it will only be noted that, in the case of notched specimens, the stress values given are the nominal values calculated assuming elastic behavior and ignoring the stress-concentration effects of the notches.

Influence in Low-Temperature Fractures

A major low-temperature fracture problem in Ni-Mo-V steels is the one associated with the ductile-to-brittle transition characteristic of all ferritic steels. In studying this problem, notched specimens are often used. Previous work [1, 2][1] has shown that when notch acuity, specimen size, and testing conditions are appropriate, the brittleness may be manifested in low-notched fracture strength. When the critical region at the root of the notch contains an unusual number of nonmetallic inclusions, even greater losses in strength can occur. The inclusions in these instances consist principally of sulfides which are segregated into localized regions. The macroscopic and microscopic appearances of these regions have been discussed previously [3]. These segregation regions are relatively small in cross section but have considerable length. A typical dimension would be 0.1 in. in diameter by four in. in length. They are randomly distributed in certain areas of the forging with the long direction generally parallel to the longitudinal direction of the forging. Fig. 1 shows the distribution of these regions observed in a Ni-Mo-V forging rejected from other considerations.

Examples of low-fracture strengths whenever the notch root intersected these regions can be illustrated by some data for the notched slow-bend type of specimens. The segregation regions, if present, were oriented such that they lie in the plane of the notch and intersected the notch root nearly perpendicularly. Thus high stresses and strains are imposed transversely across these

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not of the Society. Manuscript received at ASME Headquarters, February 24, 1959. Paper No. 59—Met-10.

[1] Numbers in brackets designate References at end of paper.

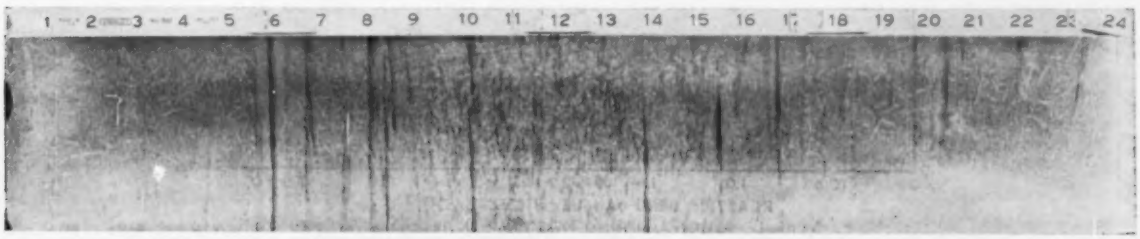


Fig. 1 Distribution of inclusion segregations in a Ni-Mo-V steel forging. Longitudinal section, etched with 10 per cent ammonium persulfate.

regions. The specimens were either 1-in. sq with a 0.2-in. deep notch finished to 0.003-in. radius or 2 in. square with 0.4-in. deep notch finished to a 0.005-in. radius. The results for several Ni-Mo-V steels tested at room temperature are shown in Table 1.

Table 1 Influence of inclusion segregation in notched-bend tests of Ni-Mo-V steels

Steel and specimen size, sq in.	Nominal bend strength, psi — No segregation at notch	Nominal bend strength, psi — Segregation at notch
A, 1	193,000	148,000
C, 2	158,000	80,000
	173,000	80,000
		117,000
E, 2	185,900	91,000

The results indicate that the strength of these severely notched specimens may be reduced by as much as one half when an inclusion segregation region coincides with the notch root. In each of the low-strength specimens, the presence of the segregation regions could be revealed in the fracture surface by macroetching the fracture. A typical appearance is shown in Fig. 2.

The probable influence of inclusions was also observed in slow-bend tests of smaller notched specimens from a Ni-Mo-V steel. The specimens tested were standard V-notch Charpy specimens. In this case, the inclusions appear mainly to influence the amount of deformation to fracture with very little effect on strength. The reason for this becomes apparent by noting the shape of the nominal stress versus deformation curve for this type of specimen shown in Fig. 3. The data shown were obtained from a number of specimens by loading in successive increments, unloading, and measuring the amount of permanent deflection which is converted to plastic bend angle from geometrical considerations. Several of the specimens were loaded to fracture and, although the amount of deformation at fracture may vary considerably, the nominal stress at fracture is little affected. Since these tests were conducted at a temperature below the fracture transition temperature of the steel, the specimens broke suddenly and completely through at fracture. Several other specimens were deformed partially to amounts indicated in Fig. 3 and then carefully examined for crack initiation in the region below the notch root. In specimens deformed to more than about 1-deg bend angle, small microcracks were found beneath the notch apex and they were usually associated with nonmetallic inclusions. A typical example is shown in Fig. 4. None of the slow-bend Charpy specimens contained large segregation regions of the kind described previously but it appears that even small inclusions can act as crack initiation points in notched specimens.

From Fig. 3 it is seen that, if a Charpy specimen contained a segregation at the notch and the deformation at fracture were reduced to a small value, the effect on nominal strength would still

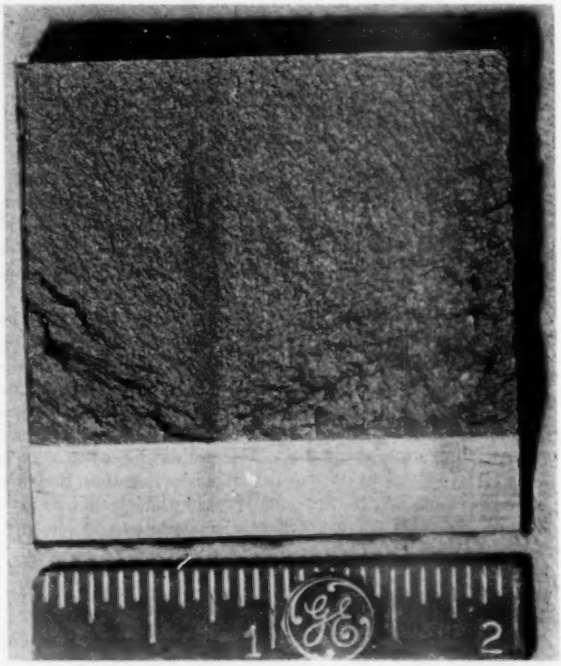


Fig. 2 Inclusion segregation in the fracture surface of a notched bend specimen of a Ni-Mo-V steel revealed by etching with 10 per cent ammonium persulfate

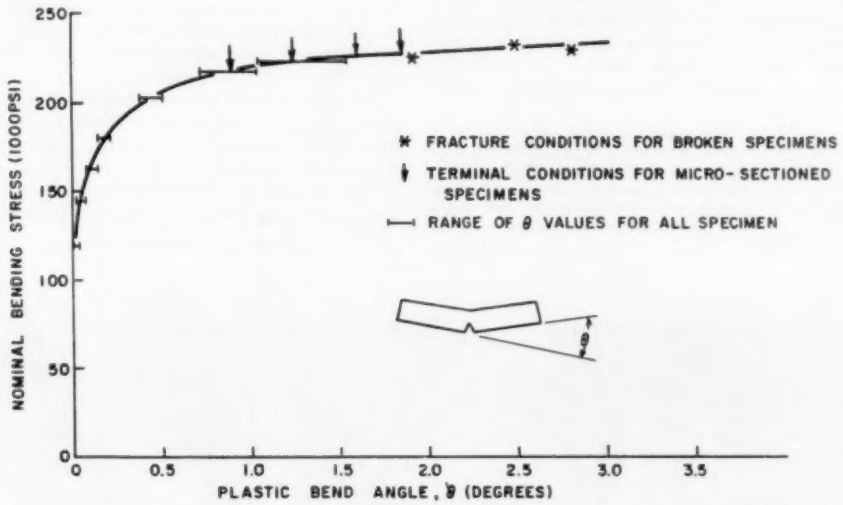


Fig. 3 Slow bend deformation and fracture behavior of V-notch Charpy specimens from a Ni-Mo-V steel, room-temperature tests

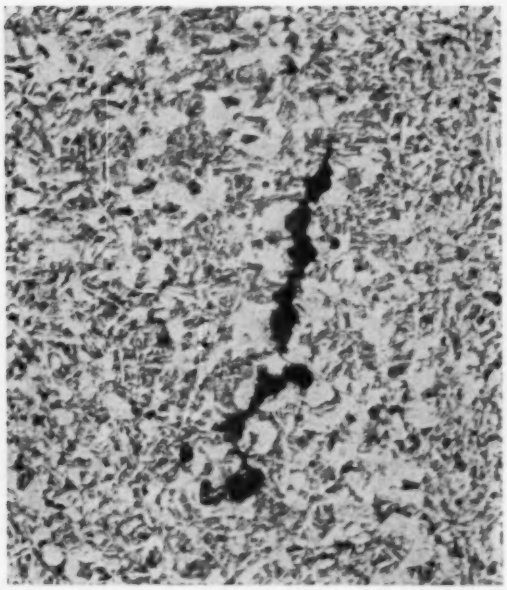


Fig. 4 A crack associated with inclusions located just beneath the notch apex in a partially deformed V-notch Charpy specimen of a Ni-Mo-V steel. Magnification, 100 X.

be small. However, the energy required for fracturing would be greatly affected since it is related to the area up to fracture under a curve such as in Fig. 3, plus a small component due to the elastic deflection of the specimen. Thus, in small notched specimens, the fracture energy values should be sensitive to the presence of inclusion segregations. A large number of V-notch Charpy impact tests had been made on the same forging as used for these slow-bend Charpy tests and the results were statistically examined. Of a total of 146 specimens from a number of locations in the forging tested at various temperatures below the fracture transition temperature, about 8 per cent had abnormally low-energy values. It was not established that every instance of a low-energy value was associated with segregation regions but it seems fairly reasonable that this would be the case. It should also be noted that this 8 per cent figure is for one forging only and the generality of the value is uncertain at present.

The microaspects of the sensitivity of these segregation regions to fracturing was also investigated by making detailed microscopic observations on small (0.05 in. thick × $^1/_8$ in. wide) prepolished unnotched bend specimens strained in small increments. In segregation-free specimens, strains of approximately 20 per cent did not result in fracture. Specimens containing a segregation region oriented so that the stress was parallel to the long direction of the segregation region did not fracture at strains of 20 per cent. However, if the segregation region was oriented transverse to the stress direction, fractures occurred at about 6 per cent strain. For either orientation of the segregate region, the larger inclusions within the segregate cracked at low strains (0.5 to 1 per cent). With the parallel orientation, the cracks in the inclusion did not extend into the matrix until large strains were imposed; while with transverse orientation, the cracks in the inclusions quickly extended into the matrix and led to final fracture. Fig. 5 shows an example of a crack in an inclusion and partially extended into the matrix. This situation probably results from the ellipsoidal shape of the individual inclusions creating the highest stress and strain concentration at its extremities.

The influence of inclusions often appears in conventional tensile tests of specimens from heavy forgings. It is a common observation that the ductilities parallel to the longitudinal direction are generally higher than in the transverse direction. Some tests indicate that inclusions which become elongated in the longitudinal direction during forging may be the predominate factor in causing this difference. Examination of specimens partially strained and also after fracture show that small fissures are formed prior to fracture in transverse specimens but not in parallel specimens. Microexamination of the fissures usually show them to be associated with nonmetallic inclusions. Apparently, either these fissures or the strain concentrations around inclusions lead to fracture at smaller over-all strain in the transverse specimens.

The association between inclusions and tensile fracture ductility may have some relationship to the correlation derived for the bursting strengths of large notched disks [4]. In this correlation, it appears that tensile ductility in addition to V-notch

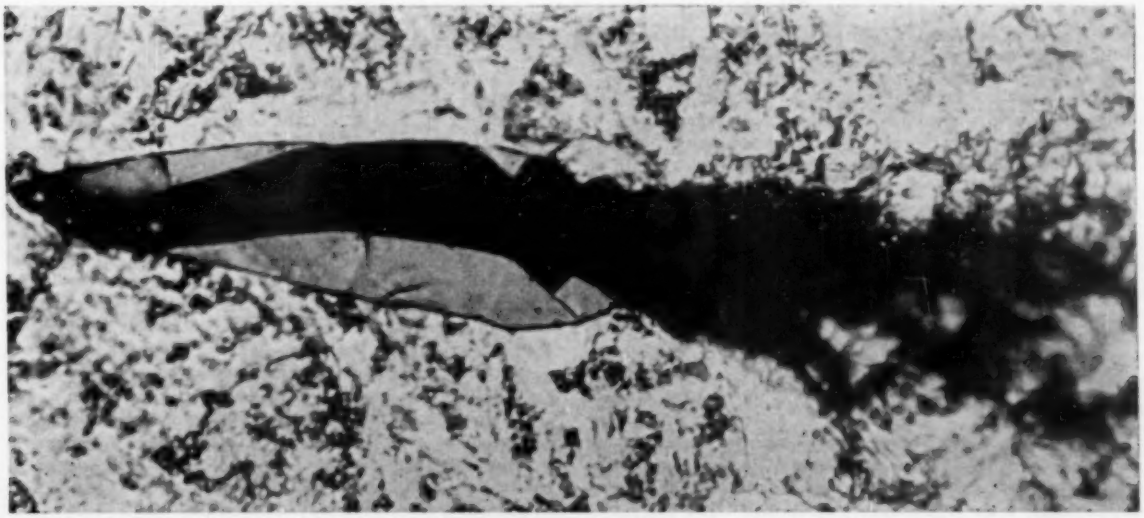


Fig. 5 A crack in an inclusion extending in the matrix in a small unnotched bend specimen of a Ni-Mo-V steel. Stressed transverse to long direction of inclusion; room-temperature test. Magnification, 250 X.

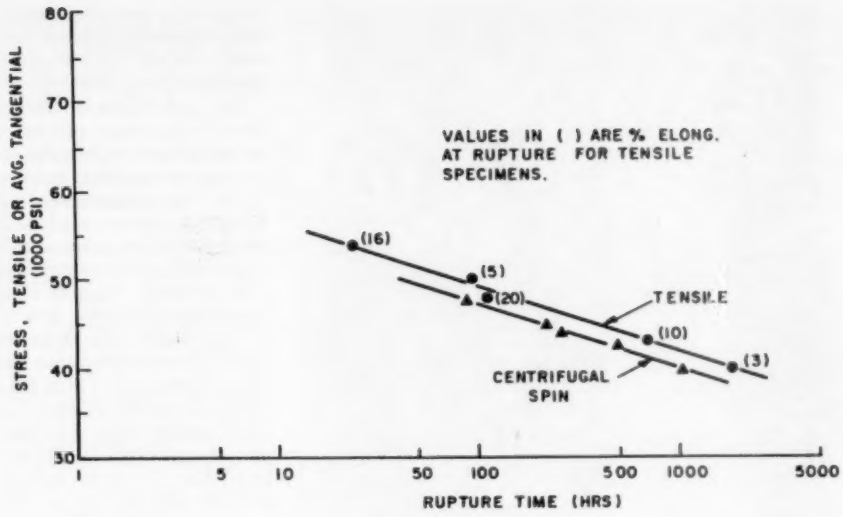


Fig. 6 Rupture life at 1000 F for tensile and disk specimens of a Cr-Mo-V steel austenitized at 1750 F

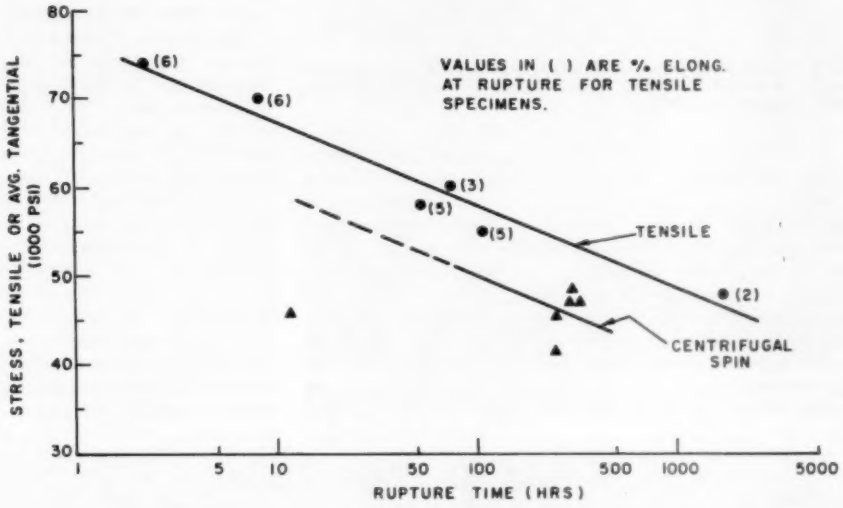


Fig. 7 Rupture life at 1000 F for tensile and disk specimens of a Cr-Mo-V steel austenitized at 1850 F

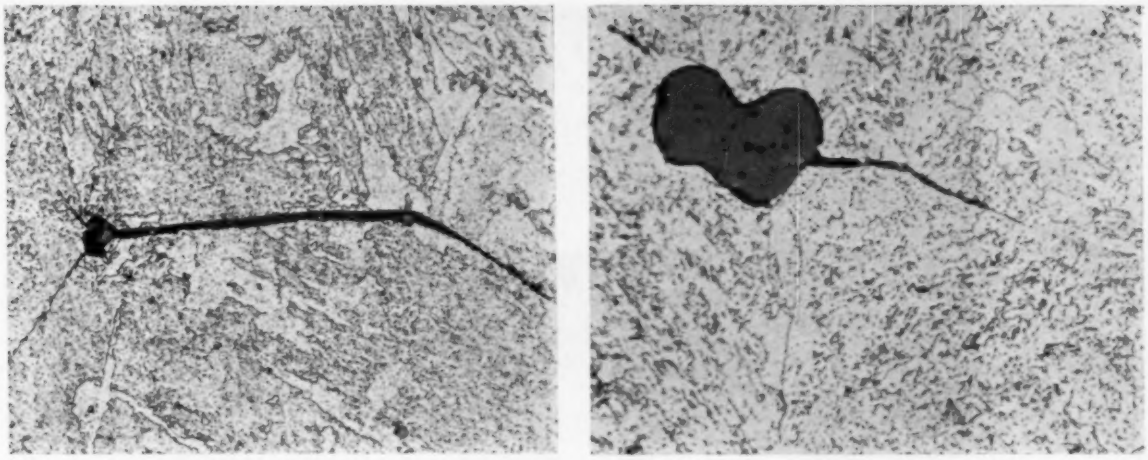
Charpy transition behavior can be a factor influencing the strength of these notched disks. It may be that tensile ductility provides an index to the existence and distribution of inclusion segregations and the possibility that they may coincide with the notches in the disk.

Influence in High-Temperature Fractures

In high-temperature strength and fracture behavior, the major problem is the formation and growth of intergranular cracks under continued stressing. Much work has been devoted to the specific influences of composition, heat-treatment, and microstructure on this phenomenon. In this section, some observations of the influence of nonmetallic inclusions in the Cr-Mo-V type of steels are presented.

These observations are for specimens from two types of tests, the unnotched tensile stress-rupture and the centrifugal spin-rupture of bored disks. The tensile specimens were of conventional design with 0.252-in-diam test section. The disk specimens were 11 in. in diam, 0.5 in. thick, with a 1.25-in-diam center hole. Specimens were obtained from two forgings which differed principally in heat-treating temperatures, in that one was austenitized at 1850 F and the other at 1750 F. The stress-rupture results for tests at 1000 F are shown in Figs. 6 and 7. The stress values for the disk specimens are in terms of an average tangential stress which is the centrifugal load divided by the net cross-sectional area along a diameter. For the material austenitized at the lower temperature, the disk-rupture times are almost equal to the tensile-rupture times. In the case of the higher austenitizing temperature, the disk-rupture data are very limited and erratic but it appears that the disk-rupture life is somewhat less than the tensile-rupture life. The material austenitized at 1850 F has a higher tensile stress-rupture strength but lower ductility and evidently, it is this deficiency in ductility which precludes a realization of the higher strength potential in a disk.

The disk-rupture tests were periodically interrupted for visual inspection for crack formation and considerable differences in behavior between materials were noted. In the specimens from the material austenitized at the lower temperature, cracks formed near the center hole at about one half of the final rupture life and continued to grow until final fracture occurred. In the specimens from the material austenitized at the higher tempera-



(a) (b)

Fig. 8 Typical intergranular cracks in intact areas of tensile and spin-rupture specimens of a Cr-Mo-V steel austenitized at 1850 F. Magnification, 250 X for (a) and 1000 X for (b).

ture, no visually observable cracks formed even on inspection at very close to the final failure time. These differences in behavior have not been completely rationalized at present but they do indicate the possibility of a significant relationship between low tensile-rupture ductility and a propensity to a sudden type of fracturing in disk-rupture tests at elevated temperatures.

It is a general observation that low tensile-rupture ductility is associated with extensive intergranular crack formation. Microscopic examination of the intact areas of both the tensile and disk specimens showed this to be the case, particularly in the specimens from the steel heat-treated at the higher austenitizing temperature. Cracks were located throughout the length of the tensile specimens but were concentrated near the center hole in disk specimen. The pertinent feature of these cracks was that, in nearly every instance, they were associated with nonmetallic inclusions as evidenced by micrographs in Fig. 8. Most of these nonmetallics seem to be either sulfides or silicates. Cracks associated with inclusions were also found in the steel austenitized at the lower temperature although, in this case, they seemed to be associated with regions having a much larger concentration of inclusions.

These results imply that increased elevated-temperature strength in nonuniformly stressed components such as rotating disks is not achieved by simply increasing the tensile stress-rupture strength of the material at the expense of rupture ductility. Furthermore, in attempting to increase the rupture ductility of steels, it appears that attention should be devoted to minimizing the amount of nonmetallic impurities. It should also be mentioned that the lower austenitizing temperature is representative of present heat-treating practices for Cr-Mo-V forgings.

Influence in Cyclic-Loading Fractures

Several investigators [5, 6] have already pointed out the generally detrimental effect of inclusions on the fatigue strength at many numbers of cycles. The improved fatigue properties of vacuum melted steels are usually attributed to this factor. The observations in this section refer specifically to influence of nonmetallics on the cyclic-loading behavior when the number of cycles to failure is relatively small and, as a consequence, the stresses are high. Furthermore, the studies deal with steels tested below their ductile-to-brittle transition temperature, so that fracturing by brittle crack propagation is also possible.

In one group of experiments, large notched bars were subjected to cyclic-bending loading. The loading was unidirectional so that the notched side goes through a zero-to-maximum tension cycle. With sharp notch radii and high nominal stresses, crack initiation occurs very early and most of the cyclic life is spent in fatigue crack propagation. The specimen fails when the local conditions at the edge of the deepening fatigue crack becomes critical with respect to initiation of a brittle, fast running fracture. The number of cycles that elapses before this critical condition occurs is determined by the load level imposed, the stress increase due to reduction in load-carrying area, the sensitivity of the steel to brittle fracturing and the existence of localized ductility deficient regions in the steel.

It has been observed in such tests on Ni-Mo-V steel specimens that, when an inclusion segregation region of the kind described previously is present in the plane of the notch and intersects the notch root, a large decrease in cyclic life occurs. In one set of comparative tests, the presence of a segregation region reduced the cyclic life by about one half. The shorter cyclic life was also accompanied by a smaller depth of fatigue crack propagation before brittle fracture occurred on the final cycle. These results for cyclic-loading conditions are mainly a consequence of the reduction in the single cycle-loading fracture strength due to the inclusion segregation. The decrease in cyclic life does not necessarily mean that a fatigue crack propagates more rapidly when a segregation is present.

Another example of the influence of inclusion segregation regions is illustrated by the series of specimens shown in Fig. 9. In these specimens, the notch axis is oriented parallel to the long direction of the segregation regions. The dark appearing spots visible on the side surface of the specimens locate the segregations. It can be seen that, with this situation, a fatigue crack which starts at the notch bottom may, or may not, encounter a segregation in the course of its propagation. For a given stress, longer cyclic life can be expected if the encounter distance is large. The specimens in Fig. 9 were 1.5 in. sq with a 0.3-in. deep notch finished to a 0.005-in. radius. When cycled to 125,000 psi maximum nominal bending stress, the cyclic lives ranged from 3783 to 5626 cycles. It is evident from Fig. 9 that increased cyclic life is associated with a longer encounter distance in these specimens.

In addition to these situations where the growing fatigue cracks may encounter inclusion segregations, the inclusions themselves

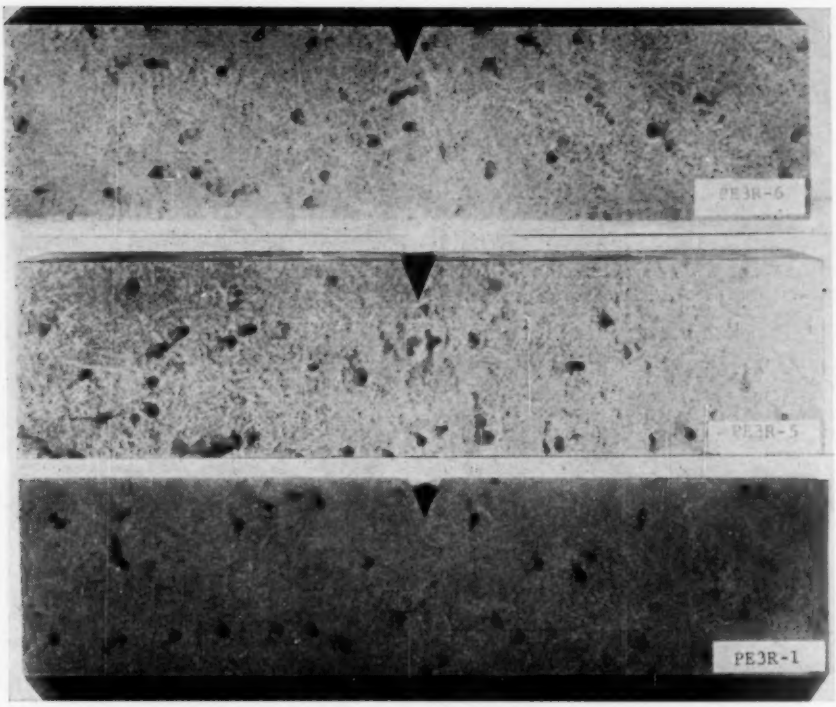


Fig. 9 Distribution of inclusion segregations in high-stress notched fatigue specimens of a Ni-Mo-V steel. Cycles to failure for identical cycling loads: Upper specimen, 3783; middle specimen, 4064, bottom specimen, 5626.

can serve as initiation points of fatigue cracks in the case of unnotched specimens. This aspect of the fatigue problem has been discussed in reference [6] in which it is pointed out that, under low-stress high-cycle conditions, a crack will be initiated only at the largest inclusion and this crack propagates to final fracture. For high-stress low-cycle conditions, cracks start at all sizes of inclusions and then join up to cause final fracture. The latter situation was observed in unnotched specimens of a Ni-Mo-V steel tested in axial tension-compression loading. Macroetching revealed that each of the specimens contained a number of inclusion segregations running transversely through the specimen at various points in the test section. The stress levels studied were in the range from slightly less than the elastic limit to well into the plastic range. In nearly all of the specimens, a fatigue crack initiated at one of the inclusion segregations and progressed to fracture indicating that these segregations can have an influence on low-cycle high-stress fatigue behavior even in unnotched specimens.

From these several examples, it seems clear that considerable increases in low-cycle fatigue behavior can be expected if these inclusion segregations can be avoided. As yet, the data are not sufficient to indicate how small and well dispersed the inclusions need to be to realize the improvement.

Summary and Conclusions

In studying the influence of nonmetallic inclusions on the strength and fracture properties, definitive results are difficult to obtain. However, the examples and observations presented have attempted to show some test conditions under which certain specific forms of inclusions can adversely affect the strength and fracture behavior of alloy steels at low and high temperatures and under cyclic loading.

Acknowledgments

The author is indebted to J. R. Low, Jr., and J. S. Erickson of the Research Laboratory for the microbend test results and Fig. 5; to A. Rubio of the Large Steam Turbine-Generator Department for the high-temperature tensile and spin-rupture data; and to H. K. Adams, Jr., of the Materials and Processes Laboratory for the results in connection with Fig. 9, all associated with the author's Company.

NOTE: When this paper was written, the author included information and expressed opinions believed to be correct and reliable. Because of the constant advance of technical knowledge, the widely differing conditions of possible specific application, and the possibility of misapplication, any application of the contents of this paper must be at the sole discretion and responsibility of the user.

References

1 J. D. Lubahn and S. Yukawa, "Size Effects in Slow Notch Bend Tests of a Nickel-Molybdenum-Vanadium Steel," Preprint 79, ASTM, 1958, presented at Annual Meeting, June, 1958.

2 S. Yukawa, "Temperature and Microstructure Dependence of Size Effects in Notched Bend Tests of Some Alloy Steels," AIME paper, to be presented at 1959 Annual Meeting.

3 C. Schabtach, E. L. Fogleman, A. W. Rankin, and D. H. Winne, "Report of the Investigation of Two Generator Rotor Fractures," TRANS. ASME, vol. 78, 1956, pp. 1567–1584.

4 D. R. DeForest, D. L. Newhouse, and B. R. Seguin, "Progress in the Development of Steam Turbine-Generator Rotor Materials," ASME Paper No. 57—A-280.

5 J. T. Ransom, "The Effect of Inclusions on the Fatigue Strength of SAE 4340 Steel," *Trans. ASM*, vol. 46, 1954, pp. 1254–1263.

6 H. N. Cummings, F. B. Stulen, and W. C. Schulte, "Relation of Inclusions to the Fatigue Properties of SAE 4340 Steel," *Trans. ASM*, vol. 49, 1957, pp. 482–512.

G. R. IRWIN
Superintendent, Mechanics Division,
U. S. Naval Research Laboratory,
Washington, D. C.

Fracture Mode Transition for a Crack Traversing a Plate

BARRING EXCEPTIONAL SITUATIONS a crack traversing a plate in tension takes a path at right angles to the direction of greatest tensile stress. When there is a substantial plastic reduction of plate thickness leading the crack the observed fracture mode is usually a sliding separation at 45 degrees to the plate surfaces sometimes termed a ductile or shear fracture. On the other hand when the contraction of thickness is relatively small the locus of the crack is usually a plane normal to the direction of greatest tensile stress and thus at 90 degrees to the plate surfaces. For this fracture mode designations such as brittle, flat-tensile, separation, and opening mode are sometimes used. Flat-tensile and brittle are employed here for this purpose as equivalent terms.

Often used with reference to fracture appearance, the terms ductile and brittle convey, in addition, the inference of a relatively large deformation work rate associated with a ductile as compared to a brittle mode of fracturing. Three major factors which contribute to a ductile-to-brittle fracture mode transition have been known for many years. These are increase of dimensional size, elevation of yield stress by temperature reduction or rapid stressing, and increase of microstructural weakening effects. Two examples of the latter are, (1) segregations and loose bonding at grain or solidification boundaries and, (2) a distribution of lattice strains or impurity atoms which favors stress relief by cleavage in place of stress relief by plastic deformation.

Two 1958 papers, References [1 and 2],[1] discussed conditions for the ductile-brittle fracture mode transition in terms of specimen thickness, yield stress, and critical fracture work rate in the flat-tensile fracture mode. Reference [1] was primarily concerned with a certain class of fracture strength measurement techniques. Reference [2] was a general review of fracture mechanics. In this paper the fracture mode transition itself is the major topic. The discussion will support and extend the remarks on fracture transition given in the previous papers. As in the current paper by B. M. Wundt [3] the crack extension force suggested by the Griffith theory is of central interest.

It has been pointed out [4, 1] that crack propagation is of practical importance primarily when it occurs through a section subjected to an average tensile load less than would be required for general plastic yielding. Under such conditions an analysis of stresses neglecting plastic strains has an obvious approximate validity. Advantage can be taken of this in several ways. By eliminating plastic strain complexities from the stress analysis one can define the crack-extension-force in a manner convenient both for calculation and measurements. One can then establish in a systematic way a procedure for using laboratory fracture test measurements to estimate the strength of structural components containing various size cracks.

The Crack Edge Stress Field

For the purpose of this discussion we need to consider only the portion of elastic stress field which encloses the advancing edge of the crack. The stress field to be considered is an approximate description of the stresses in an annular region which excludes the region of inelastic strains associated with the fracturing process but does not extend outward more than a small fraction of the length of the crack.

Using x, y and r, θ co-ordinates as shown in Fig. 1 the equations for the stresses in this region, as given in Reference [1] are

$$\sigma_y = \kappa \frac{\cos \theta/2}{\sqrt{2r}} \left[1 + \sin \frac{\theta}{2} \sin \frac{3\theta}{2}\right] \quad (1)$$

$$\sigma_x = \kappa \frac{\cos \theta/2}{\sqrt{2r}} \left[1 - \sin \frac{\theta}{2} \sin \frac{3\theta}{2}\right] - \sigma_{ox} \quad (2)$$

$$\tau_{xy} = \kappa \frac{\sin \theta/2}{\sqrt{2r}} \cos \frac{\theta}{2} \cos \frac{3\theta}{2} \quad (3)$$

$$\sigma_z = 0 \text{ (plane stress)} \quad (4)$$

$$\sigma_z = \nu (\sigma_x + \sigma_y) \quad \text{(for plane strain)} \quad (5)$$

where ν = Poisson's ratio.

The stress intensity factor κ is related to the Griffith theory strain energy release rate $\mathcal{G}$ by the following equations

$$\kappa^2 = \frac{E\mathcal{G}}{\pi} \quad \text{(for plane stress)} \quad (6)$$

$$\kappa^2 = \frac{E\mathcal{G}}{\pi(1 - \nu^2)} \quad \text{(for plane strain)} \quad (7)$$

The parameter σ_{ox} may be omitted since it affects the discussion only in minor ways and since it is, in any case, small compared to $\kappa/\sqrt{2r}$ when r is small. The equations for the principal extensional stresses in the x,y-plane then have the following simple form

$$\sigma_1 = \kappa \frac{\cos \theta/2}{\sqrt{2r}} \left(1 + \sin \frac{\theta}{2}\right) \quad (8)$$

$$\sigma_2 = \kappa \frac{\cos \theta/2}{\sqrt{2r}} \left(1 - \sin \frac{\theta}{2}\right) \quad (9)$$

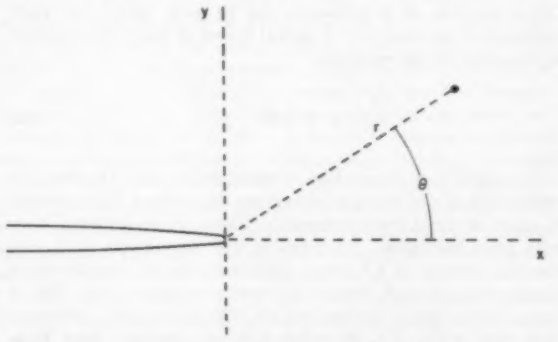


Fig. 1 Two-dimensional crack, showing the (x, y) and (r, θ) co-ordinates

[1] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, April 29, 1959. This paper was not preprinted.

when a stress in the z-direction is relevant as in a situation of plane strain the third principal extensional stress is

$$\sigma_z = 2\nu\kappa \frac{\cos \theta/2}{\sqrt{2r}} \tag{10}$$

Plane-Strain Conditions Near the Crack Edge

For purposes of mathematical analysis a brittle or flat-tensile fracture is considered to be a flat internal free surface bordered at the advancing edge by a zone of plastic strains of relatively small size. To satisfy free surface conditions at the crack surfaces, the stress system adjacent to the advancing edge must be one of plane-strain. However, for a through-thickness crack moving across a plate, the zone of yielding at the advancing edge tends to relax the stress in the plate thickness direction. Obviously a plane-strain stress sytem cannot be present where the crack intersects the plate surface. Thus a brittle fracture traversing a plate normally has a perceptible border at the plate surfaces where the mode of fracturing changes from flat tensile to shear.

Consider next the conditions under which a flat-tensile mode of fracture can be maintained in central regions of the plate thickness. An accurate representation of the stress relaxing influence of plastic strain near the advancing edge of a crack is not at present possible. However, for the general features with which we are concerned, estimates of an approximate nature can be employed. For estimating the influence of stress relaxation from a local zone of plastic strains an elastic stress distribution problem studied by Sternberg and Sadowsky [5] is helpful. For a plate in uni-axial tension Sternberg and Sadowsky were interested in how stresses near the borders of a cylindrical hole in the plate varied in character with the ratio of hole diameter to plate thickness. Their calculations indicated the stress system near the middle plane of the plate differed very little from one corresponding to a plane strain situation when the hole diameter was less than half of the plate thickness. Thus the existence of a plane strain stress situation near the advancing edge of a flat-tensile fracture in a central portion of the plate thickness suggests that the stress relaxing influence of the plastic strain zone is no more effective than that of an open round hole whose diameter is less than half the plate thickness.

From Equation (8) the distance r from the crack edge at which the predicted tensile stress equals the tensile yield stress is greatest at $\theta = 60$ deg where it has the value

$$r_Y = 0.84 \frac{\kappa^2}{\sigma_Y^2} \tag{11}$$

The diameter of the open hole whose stress relaxing influence would be comparable to that of the crack edge plastic strain zone will be taken, provisionally, as equal to r_Y. A proportionality factor adjustment, if necessary, can be made later. The plate thickness B necessary for a central region of flat-tensile fracture is then given by the equation

$$B = 1.68 \frac{\kappa^2}{\sigma_Y^2} \tag{12}$$

For application to problems of crack propagation the stress intensity values κ of primary interest are the values κ_c which pertain to onset of rapid crack extension. These values, however, vary with plate thickness. As shown in previous papers [1, 2] substantial increase of κ_c^2 occurs with reduction of plate thickness during the approach toward fracture mode transition. Fig. 2 shows this variation as observed for 7075-T6 Al alloy. The ordinate used in Fig. 2 is the value of $\mathcal{G}_c$ as computed from Equation (6). The difference from unity of the factor $(1 - \nu^2)$ of Equation (7) is neglected here regardless of the degree of plane strain.

From Equation (12) the critical value of the reciprocal of the plate thickness for fracture mode transition may be estimated as

$$\frac{1}{B} = \frac{\sigma_Y^2}{1.68\kappa_c^2} = \frac{\pi\sigma_Y^2}{1.68E\mathcal{G}_c} \tag{13}$$

A portion of the fracture transition locus line predicted by Equation (13), for $\sigma_Y = 72{,}000$ psi and $E = 10.5 \times 10^6$ psi, is shown on Fig. 2. This locus intersects the experimental $\mathcal{G}_c$ curve at the point where $1/B$ is 4.2 $(\text{in.})^{-1}$ and $\mathcal{G}_c$ is 215 lb/in. The fact that this position on the curve of Fig. 2 marks the beginning of the fracture mode transition from flat-tensile to shear is evident. A comparison of the fracture appearances for the thicknesses 0.25, 0.187, and 0.125 in. is shown in Fig. 3.

In order to compare the fracture mode transition behavior of various materials it is convenient to express plate thickness in terms of a nondimensional parameter α where

$$\alpha = \frac{B\sigma_Y^2}{E\mathcal{G}_{Ic}} \tag{14}$$

$\mathcal{G}_{Ic}$ is the value of $\mathcal{G}_c$ for completely plane-strain fracturing as shown in Fig. 2. Values of $\mathcal{G}_c$ observed for aluminum alloy plates and a rotor steel are shown in Fig. 4. Even allowing for considerable experimental scatter it is apparent that a trend curve through the observed $\mathcal{G}_c$ values would cross the Equation (13) locus line at values of α not far from unity. The fact that the fracture mode transition from flat-tensile to shear occurs with α decreasing through the range 1.5 to 0.5 will be given additional illustration later.

Abruptness of the Brittle-Ductile Transition

That the change from a brittle to a ductile fracture mode should be relatively abrupt with decrease of plate thickness can be understood as follows: In the preceding discussion the influence of the zone of yielding at the advancing edge of a crack was considered equivalent to a selected diameter hole in its tendency to relax stresses in the plate thickness direction. The equivalent hole diameter was assumed to be proportional to the size of the crack edge plastic strain zone. The shear stress corresponding in direction with the type of plastic strain which would result in plate thickness reduction is

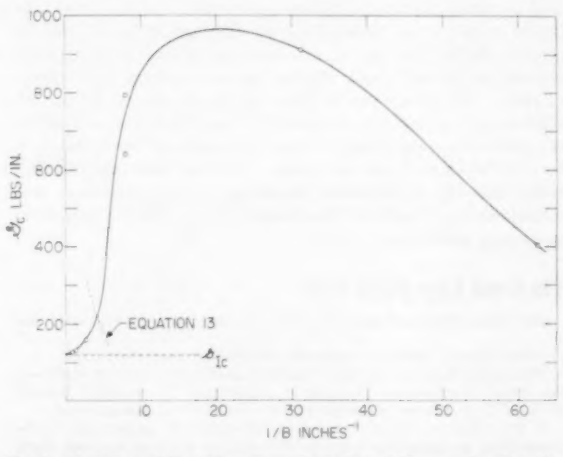


Fig. 2 Variation of $\mathcal{G}_c$ with reciprocal of plate thickness showing the $\mathcal{G}_{Ic}$ intercept

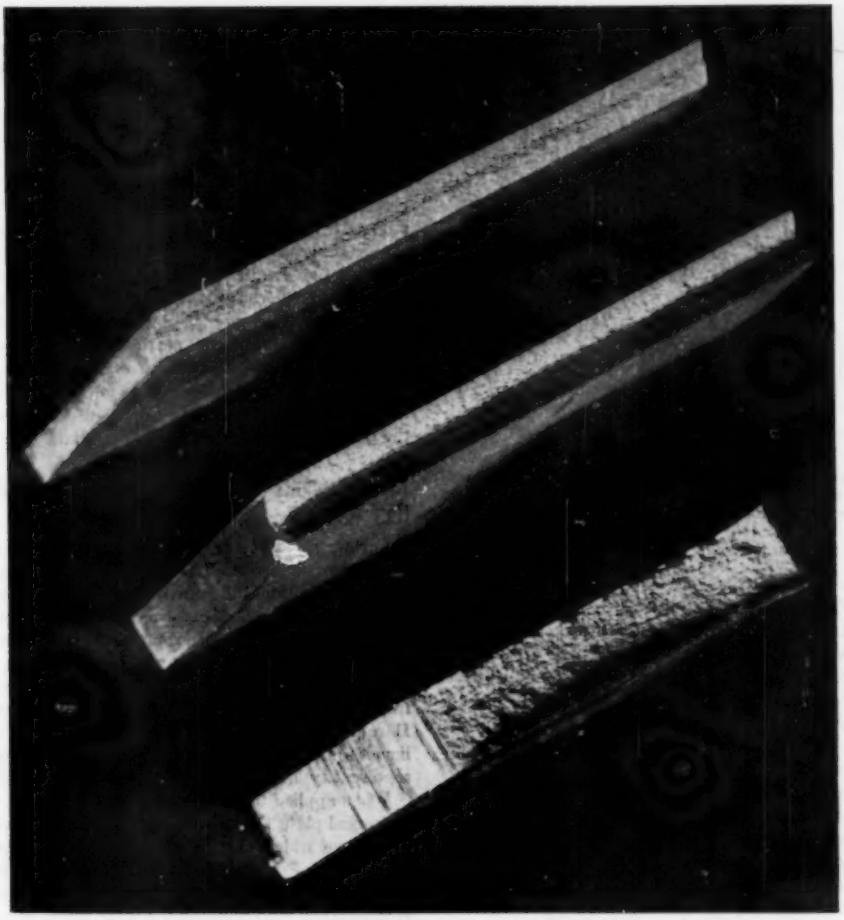


Fig. 3 Fracture appearance for 1/4, 3/16, and 1/8-in. plates of 7075-T6 aluminum broken in tension

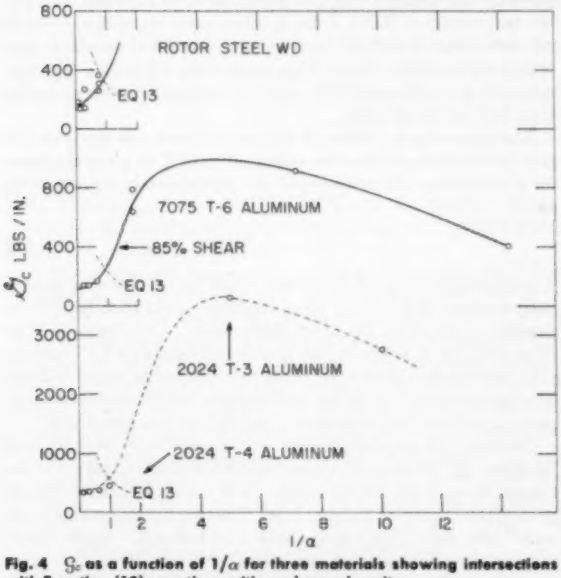


Fig. 4 $\mathcal{G}_c$ as a function of $1/\alpha$ for three materials showing intersections with Equation (13) near the position where α is unity

$$\tau_{1z} = \frac{\sigma_1 - \sigma_z}{2} \tag{15}$$

The positions r, θ from the mathematical edge of the crack where τ_{1z} has a fixed value are shown by curve A on Fig. 5 for the plane-strain stress system of Equations (8) and (10). Estimating the value of the same shear stress component by putting σ_z equal to zero to represent plane stress, the locus of the same fixed τ_{1z} value is shifted out to curve B of Fig. 5. A comparable expansion in size occurs for the crack edge plastic strain zone when a reduction of plate thickness produces a change from the flat-tensile to the shear fracture mode. When the plate thickness is too small to meet a flat-tensile fracture condition, for example Equation (12), the flat-tensile fracture mode even if initiated, say by notch sharpness, is unstable. As the crack extends, the plastic strain zone becomes larger than the size characterizing a situation of plane strain. When this occurs the shear stress component tending to cause thickness reduction increases from relaxation of σ_z. Thus the plastic zone size becomes still larger, then τ_{1z} must become larger again, and the process continues toward a situation of plane stress with complete relaxation of σ_z.

From the reasoning given the change of fracture mode with decrease of plate thickness should be abrupt requiring only a small change of plate thickness. For 7075-T6 aluminum a factor of two thickness reduction from $^3/_8$ to $^3/_{16}$ in. is required for the change

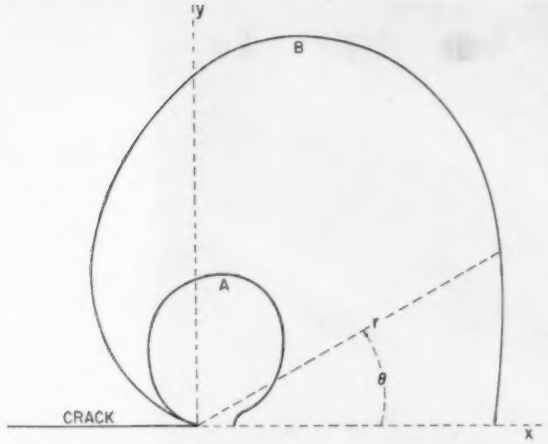


Fig. 5 Curve A shows locus of a fixed value of $(\sigma_1-\sigma_3)/2$ at the edge of a flat internal free surface in tension assuming plane strain. Curve B shows the expansion of this fixed value curve when σ_z is assumed to be zero.

from a flat-tensile fracture with small shear lips to a 90 per cent shear fracture. This may seem less abrupt than would be theoretically expected. Several factors affect the degree of abruptness of the fracture mode transition. The presence of the shear fracture side borders for a flat-tensile fracture traversing a plate causes the crack to lead in the center of the plate and lag at the edges. This assists maintaining of plane-strain stress conditions along the leading edge of the flat-tensile crack in ways not considered in Equation (12).

A second factor is present when the material has a dynamic yield stress elevation. In the usual fracturing experiment the crack extension speed increases with increase in crack length. When crack extension begins from a sharp natural crack or notch which extends through the plate thickness growth of shear lips at the side surfaces which might otherwise result in a completely shear fracture may be held to a fraction of the plate thickness if sufficient increase of yield stress occurs due to speed of crack extension. In fact, for steel plates one often observes first a growth, then a reduction, of shear lip size as a crack spreads from a sharp notch. The influence of this factor upon fracture mode transition depends upon the observed length of run of the crack and the character of the fracture test observations.

Effect of Yield Stress and Temperature

By the same reasoning as was applied to the effect of plate thickness, the crack edge plastic zone size is expected to increase with decrease of σ_Y and with increase of $\mathcal{G}_{Ic}$. These changes of σ_Y and $\mathcal{G}_{Ic}$, when of sufficient magnitude, can produce a transition from brittle to ductile fracture. For a steel plate, decrease of σ_Y and increase of $\mathcal{G}_{Ic}$ occur together with increase of testing temperature. Thus starting with conditions which result in a flat-tensile fracture, it should be possible to increase $(\mathcal{G}_{Ic}/\sigma_Y{}^2)$ enough by elevation of testing temperature to cause transition to a shear fracture. Basically the fracture mode transition occurs because the size of the plastic strain zone at the advancing edge of the crack increases relative to the plate thickness. In accordance with the line of reasoning presented for the effect of plate thickness, an abrupt change with temperature of $\mathcal{G}_{Ic}$, of σ_Y, or of both is not required for an abrupt change to occur in the mode of fracturing. The instability toward fracture transition with growth of the plastic zone size is inherent in the stress considerations previously discussed. Although a temperature dependent change in the mode of crystalline separation may exist and may have a significant influence through its effect upon $\mathcal{G}_{Ic}$, the general characteristics of a fracture transition can be understood without specific reference to the mode of crystalline separation.

Measurements published by Winne and Wundt [6] for flat-tensile fracture of two turbine rotor steels indicate a factor of two increase of $\mathcal{G}_c$ with a temperature increase from room temperature of 200 F. In the same interval $\sigma_Y{}^2$ decreased by only 14 per cent. Examples of a factor of two change in $\sigma_Y{}^2$ in a 200 F temperature interval are rare. One instance of such a change can be found in the dynamic yield stress data for a ship steel studied by Tipper, Sullivan, and Krafft [7]. Thus for a fracture mode change with temperature the change of $\mathcal{G}_{Ic}$ is usually the factor of dominating importance. However, the influence of yield stress upon the α value cannot be neglected particularly in estimating the effect of impact as compared to static stressing of a test specimen. Furthermore, the equation for α does not represent explicitly the entire influence of a change in σ_Y because $\mathcal{G}_{Ic}$ must be sensitive to the plastic flow resistance of the material.

Measurements and Calculations Related to Fracture Transition

Certain measurements on aluminum alloys and steel have already been discussed. The table of results given in Reference [1] provides additional illustrative data. In Table 1 of this paper some of the previous data for metals are repeated with the values of $\mathcal{G}_{Ic}$ recalculated using the empirical adjustment equation at the bottom of the table. Recent measurements of $\mathcal{G}_c$ for high strength sheet steels are included. The tabulated results were extracted from experimental data obtained by Herschel Smith, NRL, Harold Bernstein, NGF, and Hugh Romine, NPG. Although believed to be correct for the samples tested the steel data may not represent typical fracture toughness properties of the materials.

When applied to an experimental curve of $\mathcal{G}_c$ as a function of reciprocal plate thickness, Equation (13) indicates a position on this curve which borders the fracture mode transition on the brittle side. Primarily Equation (13) expresses the fact that only a limited size zone of plastic yielding at the leading edge of a crack is compatible with maintaining plane strain conditions in central regions of the plate. The value of α pertaining to this point is apparently not far from unity for materials so far investigated. For the test results of Table 1 the fracture mode transition, from 40 per cent shear to 100 per cent shear, is completed for all the materials tested within the α-value range from 1.2 to 0.5. The corresponding α-value range for a specific material appears to be less than half of this interval.

The introduction and use of the parameter α was motivated in part by the thought that the relationship of $\mathcal{G}_c$ to plate thickness for a wide range of metals might be approximately representable as

$$\mathcal{G}_c = \mathcal{G}_{Ic} f(\alpha) \qquad (16$$

A valid relationship of this form would be quite useful since a measurement of $\mathcal{G}_c$ at one plate thickness could then be used to predict values of $\mathcal{G}_c$ for other thicknesses. Data such as that shown on Fig. 4, although not closely representable by Equation (16), nevertheless show a sufficient resemblance in major features to suggest ways in which the relative plate thickness parameter α, can be useful in the analysis of fracture test measurements.

Consider, for example, recent measurements by Srawley and Beachem [8] of fracture appearance transition temperature for certain sheet steels having tensile yield strengths above 180,000 psi and thicknesses of $^1/_8$ in. or less. The test specimens employed were $^1/_4$-in. wide tensile strips with a central slot. Brittle crack extensions at both ends of the central slot were developed prior to testing. The fracture mode transition produced in these tests

Table 1

Thickness, in.	Aluminum alloys	σ_y, 10^3 psi	$\mathcal{G}_c$[a], lb/in.	G_{Ic}[b],	α	Per cent shear
1.00	7075-T6	72(Nom.)	130	115	4.30	...[c]
0.75	7075-T6	72	134	115	3.22	...
0.367	7075-T6	72	158	115	1.57	...
0.187	7075-T6	72	...	115	0.80	85
0.125	7075-T6	72	639	115	0.54	100
0.125	7075-T6	72	792	115	0.54	100
0.032	7075-T6	72	912	115	0.137	100
0.016	7075-T6	72	403	115	0.069	100
8.00	2024-T4	50(Nom.)	327	300	6.4	...
4.00	2024-T4	50	347	300	3.2	...
2.00	2024-T4	50	367	300	1.6	...
1.00	2024-T4	50	449	300	1.0	...
0.25	2024-T3	50	4260	300	0.20	100
0.125	2024-T3	50	2740	300	0.10	100
	Steels					
0.250	SAE 4340	230	140	125	3.5	Small
0.075	SAE 6150	250	275	134	1.16	50
0.084	Tricent	200	290	116	0.97	50
0.095	X-200	242	164	97	1.91	10
0.072	X-200	240	230	116	1.19	15
0.072	X-200	194	1030	174	0.52	80
0.130	Vasco X-3, G′	225(Nom.)	84	68	3.4	Small
0.085	Vasco X-3, G′	225(Nom.)	173	104	1.34	15
	AMS 6434(0.36 C)					
0.080	AT-3,[d] 4-400°	210	338	126	0.93	45
0.080	AL-5,[d] 6-400°	219	665	178	0.71	75
0.080	AT-9, 10, 11, 12-500°	213	261	115	1.05	40
0.080	AL-15, 16-500°	216	1130	217	0.57	95
0.080	AT-1, 2-400°	210	290	120	0.98	40
0.080	AL-7, 8-400°	219	650	175	0.73	60
0.080	AL-13, 14-500°	216	1330	227	0.55	85
0.080	AT-17, 18, 19-725°	211	257	113	1.05	50
0.070	AT-20-725°	211	279	109	0.95	60
0.080	AL-21, 22-725°	207	1130	204	0.56	95
0.195	Longitudinal-RT	215	771	292	0.97	60
0.195	Longitudinal-0°F	(223)	652	297	1.09	60
0.196	Transverse-RT	214	722	300	1.00	50
0.196	Transverse-0°F	(221)	634	291	1.10	50
	Titanium B-210 VCA					
0.090	Longitudinal	149	1600	264	0.53	85
0.100	Transverse	146	910	222	0.67	70
0.091	Longitudinal	165	700	218	0.79	60
0.097	Transverse	161	280	125	1.40	10
0.092	Longitudinal	178	265	151	1.38	5
0.097	Transverse	177	175	122	1.78	<5

[a] Corrected for plastic zone stress relaxation.

[b] Computed by assuming $\mathcal{G}_c = \mathcal{G}_{Ic}\left(1 + \frac{1.4}{\alpha^2}\right)$.

[c] Not estimated unless crack length was more than three times plate thickness.

[d] AL and longitudinal specimens were broken across the direction of stringer orientation. AT and transverse specimens were broken along the direction of stringer orientation.

by variations of the testing temperature was sharply characterized by the point at which the shear lips spread together giving a 100 per cent shear fracture.

Fig. 6 from the Srawley-Beachem data shows the critical plate thickness for the 100 per cent transition as a function of temperature for two steels. Steel B was a 12 Cr, 0.25 C martensitic stainless steel somewhat embrittled at the 900 F tempering temperature. Steel G was a hot-work die steel. Both steels and their heat-treatments are described in Reference [8]. The dashed lines in Fig. 6 are considered equivalent to a 0.2 per cent offset tensile yield stress. However, precise yield stress measurement was not of central interest in this work. For steel B a single yield stress-temperature relation is shown for all three sheet thicknesses. For steel G the measurements suggested the yield stress of the 0.10-in. sheet exceeded that of the two thinner sheets through out the range of testing temperatures.

Using Equation (14) and the Srawley-Beachem curves for fracture appearance as a function of temperature, values of $\mathcal{G}_{Ic}$ were calculated on the assumptions that α had the fixed value 1.0 for the 50 per cent shear fracture point and the fixed value 0.7 for the 100 per cent shear fracture point. In view of the steepness of the curves in this region it was evident this portion of the fracture mode transition corresponded to a smaller change of α. However, it was felt the results of the computations might serve both to bracket the range of probable $\mathcal{G}_{Ic}$ values and at the same time would indicate the uncertainty in estimates of this type. The results are shown in Fig. 7. Also shown in Fig. 7 are measurements of $\mathcal{G}_c$ from centrally notched spin-disk tests reported by Winne and Wundt [6]. The selected spin-disk results are for a low fracture strength rotor steel. Because the 6-in. thick disks were equivalent to large values of α, the correction of these results to $\mathcal{G}_{Ic}$ values would merely lower all of them by a small constant fraction, about 30 per cent, and would not change the trend of the results with testing temperature.

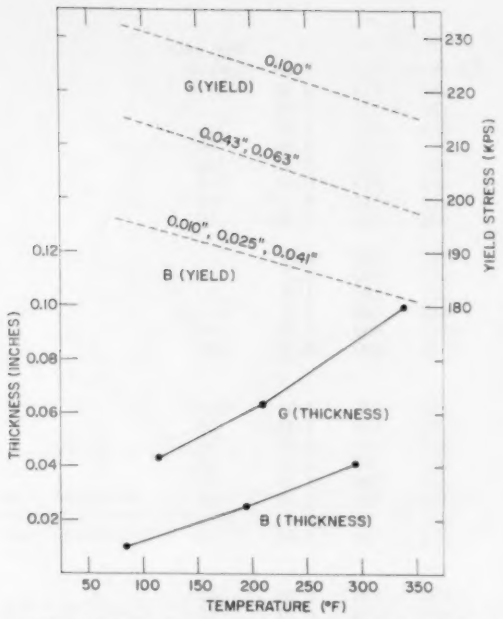


Fig. 6 Critical thickness for transition to 100 per cent shear fracture as a function of temperature for two steels. Upper curves show tensile yield stress. (Data from Srawley and Beachem.)

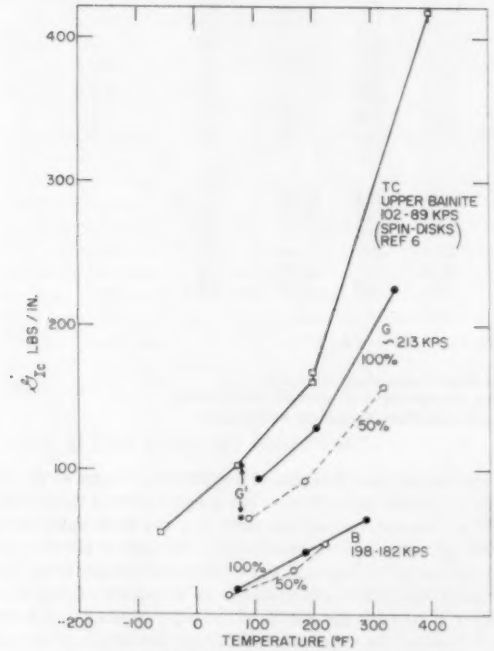


Fig. 7 Trend of $\mathcal{G}_{Ic}$ with temperature for several steels

The trends with temperature of the values of $\mathcal{G}_{Ic}$ calculated from Equation (14) are quite similar for steels B and G. From the close correspondence of these results to the trend of $\mathcal{G}_{Ic}$ with temperature for rotor steel TC it is believed estimates of $\mathcal{G}_{Ic}$ from Equation (14) combined with observations of fracture mode transition will be of considerable interest.

Measurements of $\mathcal{G}_c$ were made on two thicknesses of Vascojet X-3 which is similar to steel G. Yield stress was not measured. From the hardness and experience with similar materials the tensile yield stress should be close to 225,000 psi. The results of these tests in terms of values of $\mathcal{G}_{Ic}$ are indicated in Fig. 7 by the symbol G′. One set of test plates were thick enough, 0.130 in., so that the adjustment, using the equation under Table 1, from $\mathcal{G}_c$ to $\mathcal{G}_{Ic}$ was small. The Table 1 adjustment equation is an arbitrary choice of Equation (16) which fits the low $\mathcal{G}_c$ region of the fracture transition for 7075-T6, a low strength rotor steel studied by Winne and Wundt, and a rotor steel studied by Wessel. Although the range of validity of this $\mathcal{G}_c$-to-$\mathcal{G}_{Ic}$ adjustment equation is not well established, the α values in Table 1 nevertheless appear to be in fair agreement with the degree of fracture mode transition as indicated by per cent shear.

The accuracy with which one can calculate $\mathcal{G}_{Ic}$ from observations of the fracture mode transition either with change of plate thickness or change of temperature depends upon the accuracy with which a selected feature of the fracture transition can be characterized by a value of the parameter α. From available data the change of fracture appearance from 40 per cent shear to 95 per cent shear does not span more than a factor of two change of α value. Thus the error of the α estimate should be substantially less than a factor of two. With some refinements an accuracy comparable to that of $\mathcal{G}_c$ type fracture strength measurements, say 10 to 20 per cent, might be realized.

In the case of steels of low hardness and large dynamic increase of yield stress direct experimental measurement of $\mathcal{G}_{Ic}$ is difficult, particularly for dynamic loading conditions, and may have no more accuracy at present than the indirect procedure using Equation (14) together with observation of fracture transition and dynamic yield stress. The Equation (14) procedure also merits consideration for estimating the flat-tensile fracture strengths of high yield strength sheet steels used in construction of rocket engine chambers. Direct measurements of $\mathcal{G}_{Ic}$ for these steels are inconvenient because of the small sheet thickness. However, the $\mathcal{G}_{Ic}$ values are of interest. The spread of a surface crack through the sheet thickness under plane strain conditions has been noted as the starting phase in a number of instances of fracture failure of solid propellant engine chambers. Therefore both as an aid to inspection and to fracture failure analysis it is desirable to know the crack size which will propagate through the sheet thickness under various tensile stress loads.

Conclusions

The usually observed abrupt change from brittle to ductile fracture with increase of temperature as well as the less commonly studied change of fracture mode with plate thickness can be ascribed to the unstable growth tendency of the plastic zone at the leading edge of the crack when conditions necessary for a crack edge stress environment of plane strain in central portions of the plate thickness are not present. From this viewpoint the fracture mode transition is controlled by a relative plate thickness parameter α which measures approximately the ratio of the section thickness to the size of the crack edge plastic strain zone. Knowledge of values of α characteristic of selected readily observed features of the fracture within the region of a fracture mode transition permits an estimate of the value of $\mathcal{G}_c$ for completely plane strain fracturing.

References

1 G. R. Irwin, J. A. Kies, and H. L. Smith, "Fracture Strength Relative to Onset and Arrest of Crack Propagation," *Proceedings, ASTM*, vol. 58, 1958, p. 640.

2 G. R. Irwin, "Fracture Mechanics," First Symposium on Naval Structural Mechanics, August 11–14, 1958 (Stanford Univ.).

3 B. M. Wundt, "A Unified Interpretation of Room Temperature Strength of Notched Specimens as Influenced by Their Size," ASME Metals Engineering Division Conference, Albany, N. Y., April 29–May 1, 1959.

4 M. E. Shank, "A Critical Survey of Brittle Failure in Carbon Plate Steel Structures Other Than Ships," ASTM STP No. 158, pp. 45–110, 1953.

5 E. Sternberg and M. A. Sadowsky, "Three-Dimensional Solution for the Stress Concentration Around a Circular Hole in a Plate of Arbitrary Thickness," *Journal of Applied Mechanics*, vol. 16, TRANS. ASME, vol. 71, 1949, pp. 27–38.

6 D. H. Winnie and B. M. Wundt, "Application of the Griffith-Irwin Theory of Crack Propagation to the Bursting Behavior of Disks, Including Analytical and Experimental Studies," TRANS. ASME, vol. 80, 1958, pp. 1643–1655.

7 J. M. Krafft, A. M. Sullivan, and C. F. Tipper, "The Effect of Static and Dynamic Loading and Temperature on the Yield Stress of Iron and Mild Steel in Compression," *Proceedings, Royal Society*, series A, vol. 221, 1953, pp. 114–127.

8 J. E. Srawley and C. D. Beachem, "Crack Propagation Tests of Some High Strength Sheet Steels," NRL Report 5263 Jan. 10, 1959.

DISCUSSION

Frank A. McClintock[2]

Dr. Irwin has shown clearly how the transition from a plane strain to a plane-stress mode of fracturing can be related to the ratio of the plate thickness B to the critical radius of the plastic zone for the onset of crack instability under plane-strain conditions R_{Ic}, as indicated by his Equation (16). But as he noted, tests at one thickness do not give the notched strength at all thicknesses. This is not surprising, for where there is a change in the metallurgical mode of fracture as in the case of mild steel, for example, one would not expect the relative notched strength in the two regions to bear the same relation as they do for materials such as the aluminum alloys which do not seem to have a metallurgical fracture transition. Thus the plane-stress and plane-strain regions should be considered separately in trying to estimate the notched strength from more directly observable properties, say, by analogy with the instability of shear cracks for which an analysis is available [9].[3] That analysis dealt with an elastic, perfectly-plastic material of yield strain in shear γ_y, which was assumed to fracture when a strain γ_f was attained throughout a region of size ρ, within which the plastic deformation could not be considered homogeneous. The critical radius of the plastic zone for instability of the crack R_c was given approximately by

$$R_c = \rho \exp\left[(2\,\gamma_f/\gamma_y) - 1\right]^{1/2} \tag{17}$$

The corresponding value of the nominal stress was given in terms of the yield stress in shear k, the crack length c, and the radius of the plastic zone R, by the approximate equation

$$\frac{\tau}{\kappa} = \left\{(1 + R/c)\left[\sqrt{1 + \frac{4(R/c)\left(\frac{4}{\pi} - 1\right)}{\left(1 + \frac{R}{c}\right)^2}} - 1\right] \Big/ 2\left(\frac{4}{\pi} - 1\right)\right\}^{1/2} \tag{18}$$

At low stress levels this relation, plotted in Fig. 8 of this discussion, approaches the familiar dependence on the reciprocal of the square root of the crack length.

To establish an analogy between the plane stress and the shear solutions, note that in the plane-stress case one would expect to find elastic, plastic, and necking regions such as shown in Fig. 9. In the plastic region one might assume that there was just enough workhardening to offset the decrease in thickness until some critical value of the strain was reached, at which point necking would begin. This critical value of the strain can be termed the uniform strain ϵ_u and, while it depends on the ratios of the principal stresses, as a first approximation it might be taken to be the uniform strain in a tensile test on a thin sheet. Since the necking of thin sheet contributes little to its over-all elongation, the uniform strain may be taken to be the elongation of sheet material in the ordinary tensile test. In order to permit the necking to go to completion and form a crack it would be necessary for necking to occur over an area of the order of the thickness of the specimen. Since the necked region could not be handled with the plane-stress equations used in the rest of the plastic zone, it might be assumed that necking would be possible if the equations indicated that the uniform strain ϵ_u was exceeded over the area which was expected to neck, namely, a region whose size is approximately the thickness of the specimen. Since complete necking is not necessary before a shear fracture, a still better approximation might be to require that the uniform strain be exceeded in a region whose size is the thickness of the specimen multiplied by the reduction of area at fracture in sheet specimens in which the necking was constrained to occur in a direction normal to the axis of maximum principal stress, as for example Fig. 10.

Finally, in approximating a material which strain hardens, the analog of the yield stress in shear should be the tensile strength of a flat strip and the analog of this yield strain should be the tensile strength divided by Young's modulus. Thus the quantities γ_f, ρ, and γ_y become

$$\gamma_f \to \epsilon_u, \qquad \rho \to B(\text{R.A.}), \qquad \gamma_y \to \text{T.S.}/E \tag{19}$$

When these equations are substituted into Equations (17) and (18) for the case when the crack length is long enough so that instability occurs at a small fraction of the tensile strength, the following equation is obtained

$$\sigma = \text{T.S.}\sqrt{\frac{B(\text{R.A.})\exp\left[(2\epsilon_u E/\text{T.S.})^{1/2} - 1\right]}{c}} \tag{20}$$

For very thin foils the structural-size parameter obtained from the foil thickness times the reduction of area may become smaller than a structural-size parameter set by the grain structure or other properties of the material itself [9]. In this case the structural size is taken from observed inhomogeneities in the deformation of the material rather than from the foil thickness.

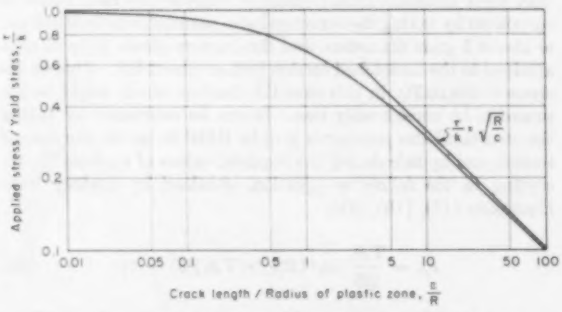


Fig. 8 Dependence of radius of plastic zone on applied stress

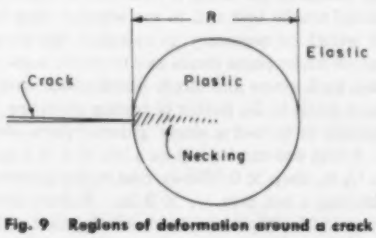


Fig. 9 Regions of deformation around a crack

[2] Professor, Massachusetts Institute of Technology, Cambridge, Mass. Assoc. Mem. ASME.

[3] Numbers in brackets designate additional References at end of this discussion.

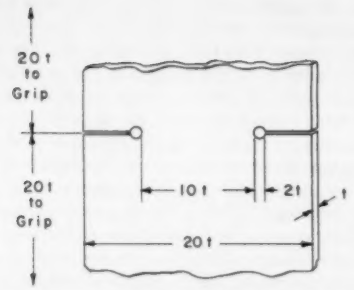


Fig. 10 Tensile specimen to produce necking normal to axis of tension

In order to check the validity of Equation (20), tensile tests were run on strips of 7075 T6 and 2024 T3 sheet, both straight and constrained to neck normal to the axis as shown in Fig. 10. For comparison with these results, the stress for instability was also calculated from Fig. 4, assuming that $\mathcal{G}_e/\alpha$ is constant for specimens as thin as, or thinner than, the thinnest indicated there, and from the equation

$$\sigma = \sqrt{E\mathcal{G}_e/\pi c} = \sqrt{E\alpha(\mathcal{G}_e/\alpha)/\pi c} \tag{21}$$

$$= \sigma_y \sqrt{\frac{B(\mathcal{G}_e/\alpha)/\pi\mathcal{G}_{Ic}}{c}}$$

The results are summarized in Table 2 and, while the agreement is by no means exact, it does appear good enough to suggest similar calculations and experiments with other materials, as well as a more detailed analysis of the elastic-plastic plane-stress problem.

For very thick plates the problem approaches that of fracture under plane-strain conditions. In the limiting case the thickness of the plate is itself no longer important, as the author has pointed out, and the structural-size parameter ρ should be associated with the structure of the material itself. In work on notch sensitivity of 7075 T6 under torsion, Mackenzie [10] found that crack initiation from notches of various root radii could be correlated by taking the structural-size parameter to be 0.010 in., or about 5 grain diameters, and the fracture strain γ_f to be that attained in the unnotched torsion test, or about 0.7. Due to increased triaxiality in this case the fracture strain might be assumed to be considerably less. It can be estimated by taking the structural-size parameter ρ to be 0.010 in. as in the case of torsion and by calculating the required values of ϵ_{fI} from $\mathcal{G}_{Ic}$ according to the following equation, obtained by analogy from Equations (17), (18), (21)

$$\epsilon_{fI} = \frac{\text{T.S.}}{2E}\left[\ln^2\left(E\mathcal{G}_e/\pi(\text{T.S.})^2\rho\right) + 1\right]. \tag{22}$$

It is found that ϵ_{fI} is 0.03 for the 2024 T3 and 0.016 for the 7075 T6. These values of the strain are far below those encountered in the uniaxial tensile test and to see whether they have a real meaning it would be necessary to calculate the state of stress around a notch under plane strain and to devise some convenient test in which such stress and strain distributions could be simulated without going to the bother of testing an entire plate. Another possibility is to test a singly grooved plane-strain tensile specimen. A test was carried out on a bar of 1 × 3-in. cross section with a $^1/_2$-in. deep × 0.0006-in. root radius groove cut across the face, leaving a net area $^1/_2$ × 3 in. Failure occurred at a nominal stress of 25,600 psi which, considering the tensile strength of 86,000 psi in the form of a $^1/_2$-in-diam tensile specimen and taking a structural-size parameter to be 0.010 in., gave a value of fracture strain ϵ_{fI} from the analog of Equations (17) and (18) of 0.012. The agreement between this and the value of 0.016 obtained from the paper may be coincidental but does bear looking into. In particular, as the fully plastic-stress level is approached, the load-carrying capacity of a singly grooved specimen approaches that of a plane-strain specimen in the fully ductile case, whereas the load-carrying capacity of a doubly grooved specimen approaches $(1 + \pi/2)$ times that value. Thus the analog of Equation (18) may not hold at high stress levels. In addition, a certain amount of bending was present in the singly grooved bar, which was pulled between grips 68 in. apart. The bending effect varies with stress level.

Table 2

Material	2024 T3	7075 T6
Thickness, B	0.125	0.050
Uniform strain in unnotched strip, ϵ_u	0.17	0.13
Reduction of area in constrained strip, R. A.	0.06	0.17
Tensile strength of unnotched strip, T. S., psi.	70,000	83,000
Modulus of elasticity, E, psi	10^7	10^7
$\sigma\sqrt{c/B}$ from Equation (20)	340,000	340,000
$\sigma\sqrt{c/B}$ from author's data and Equation (21)	270,000	290,000

If the plane-strain and the plane-stress regimes of operation can be determined from simple tests on the material as described in the foregoing then the transition between these regimes may be described in a way similar to that proposed by the author, except that two parameters, the structural size ρ and the fracture strain ϵ_f, must be described as the specimen thickness increases relative to the radius of the plastic zone. When the thickness of the specimen is very great compared to the radius of the plastic zone, the plastic region will prevent contraction and local necking. As the specimen is thinned down to the point where the elastic region ahead of the plastic zone begins to offer less constraint, one might at first think that the unyielding material either side of the originally cut notch would itself prevent contraction and

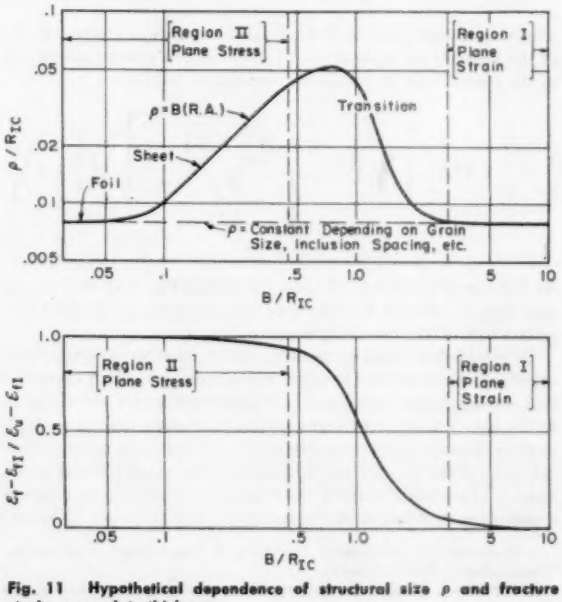


Fig. 11 Hypothetical dependence of structural size ρ and fracture strain ϵ_f on plate thickness

maintain a plane-strain situation. But as the crack grows, stably at first, this original undeformed material is left behind and necking can occur as readily as it would in the usual tensile test of a straight strip. For materials where fracture occurs before necking, there would be no change in the structural-size parameter from the plane-stress to the plane-strain condition. For materials which do neck, however, the transition of the parameters for different materials might lie on curves such as those shown in Fig. 11. For convenience and ready reference the plate thickness is nondimensionalized in terms of the radius of the plastic zone in the plane-strain condition R_{I_c} found from the analog of Equation (18)

$$R_{I_c} = \alpha(\sigma_{I_c}/\text{T.S.})^2 \qquad \text{for} \qquad (\sigma_{I_c}/\text{T.S.}) \ll 1. \tag{23}$$

The thickness ratio is related to the author's thickness parameter by

$$B/R_{I_c} = \pi\alpha. \tag{24}$$

The relations discussed in the foregoing give a more general form for the dependence of notched strength on thickness than given by Equation (16). That equation, however, set the stage for this discussion, which it is hoped will provide a guide for further work on the notch sensitivity of sheets and plates.

References

9 F. A. McClintock, "Ductile Fracture Instability in Shear," *Journal of Applied Mechanics*, vol. 25, TRANS. ASME, vol. 80, 1958, pp. 582–588.

10 A. C. Mackenzie, "Crack Initiation in Circumferentially Notched Bars in Torsion," MS thesis, M.I.T., 1958.

Author's Closure

The author is indebted to the discusser for his thoughtful comments. The paper was intended to be essentially descriptive rather than theoretical. In this as well as in previous papers the author's working plan has been to develop procedures for giving quantitative characterization to major descriptive features of fracturing. The mathematical model of crack-extension instability established by McClintock [9] should possess a degree of correspondence to these major features particularly to the relative plastic zone size. However, the mathematical model is a substantial oversimplification of the actual instability process. Thus the task of establishing correspondence to the observations reported in the paper required careful consideration by the discusser. The relationship of certain features of the paper to the problem solved in reference [9] was skillfully indicated by the discusser. Although the complex problem of crack-extension instability remains in several respects beyond reach of current analytical theory, the degree of correspondence between the McClintock's mathematical model and observations reported in the paper suggests McClintock's analysis is a useful first step toward a more complete theoretical analysis of the general problem.

A. VINCKIER
Research Engineer, Metals Joining Section, Metallurgy Department, Research Laboratories, Westinghouse Electric Corporation, Pittsburgh, Pa.

R. DECHAENE
Research Engineer, Commission pour l'Etude de la Construction Metallique, University of Ghent, Ghent, Belgium

Use of the Moiré Effect to Measure Plastic Strains

A method is described which should prove useful in the study of strain distribution in a metallic member during plastic deformation. Regularly spaced lines are drawn on a test specimen of interest and on a glass plate. After exposure to strain the part is photographed through the grid of lines on the glass. The strain distortion on the part results in an interference pattern from which the amount of plastic deformation may be calculated.

I Introduction

PLANE PLASTIC DEFORMATION in metals can be measured from the corresponding Moiré effect: If two screens of slightly crossed equidistant lines are superimposed, interference patterns, called Moiré fringes, appear where the lines of the first screen intersect those of the second. Shepard and Weller [1][1] used this method to study deflections in transparent beams. J. D. C. Crisp [2] measured elongation in transparent tensile test specimens with the aid of a special Moiré interferometer. Ligtenberg [3] determined the bending moments in plates with the aid of Moiré pattern obtained by superimposing, on the same negative, the reflected image of a ruled screen which the model rejected before and after a suitable load was applied. A work by Thiry and Dantu [4] on the application of the Moiré effect in the study of plastic deformation has recently come to the notice of the authors.

If one draws fine, regularly spaced lines on an undeformed metal test specimen and also on a transparent screen and then deforms the testpiece into the plastic range, Moiré fringes are formed when the transparent master grid is superimposed upon the deformed grid. This provides the basis from which the amount of plastic deformation may be calculated. This method has the advantage over the photogrid method in that the measurements are not so tedious, that the regions of particular interest are clearly shown, and that pictures for measurement can quickly be taken at any stage of the plastic flow. It should be possible to measure elastic as well as plastic strains by using sufficiently fine screens.

II The Grid

A first attempt to obtain a grid having regular spacings of about 0.2 mm was made by drawing lines directly on the polished surfaces of a mild-steel test specimen with a sharp chisel. A similar set of lines was scribed on a sheet of plexiglas. It was found that this method was not suitable for accurate work because of irregular spacing and small line width (0.05 mm) which resulted in irregular and poorly visible Moiré fringes.

Better results were obtained by photoengraving the lines on the metal specimen. This technique has been described by J. A. Miller [5]. A 6 × 6-in. Ronchi ruling negative with 133 lines per inch was successfully used. This produced consistently clean-cut lines, with a spacing error of about 0.2 per cent of the nominal value. Superimposing the original negative on the deformed photoengraved specimen provided sharp Moiré fringes which were easily photographed and analyzed.

III Strain Analysis

The photoengraved specimen is deformed under load in the plastic range as desired. At each incremental deformation step, pictures can be taken of the Moiré fringes after superimposing the master screen on the specimen, with the lines of the screen parallel to the original direction of the lines on the undeformed specimen. Strain measurements can then be made from the prints. To do this, a convenient co-ordinate system is chosen and co-ordinate lines are drawn on the pictures. The distance between the interference fringes are measured along these co-ordinate lines in both directions. The strains and rotations of the deformed body are then derived as described for the following three conditions:

(a) Strain perpendicular to the screen lines. A model, strained uniformly in tension in a direction perpendicular to its screen lines, with the master screen superimposed onto it without rotation, is shown in Fig. 1.

To obtain the necessary interference fringes, every nth line of the undeformed master screen must interfere with every $(n-1)$th line of the elongated specimen grid or with every $(n+1)$th line of the compressed specimen grid. If we call a the distance between two neighboring screen lines, and d the distance between two neighboring interference fringes, Fig. 1, the conventional elongation e is then expressed by

$$na = d = (n-1)a(1+e)$$

hence

$$e = \frac{a}{d-a}$$

In the case of compression

$$e = -\frac{a}{d+a}$$

For small strains ($a \ll d$) this can be written:

$$e = \pm\frac{a}{d}$$

[1] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 24, 1959. Paper No. 59—Met-7.

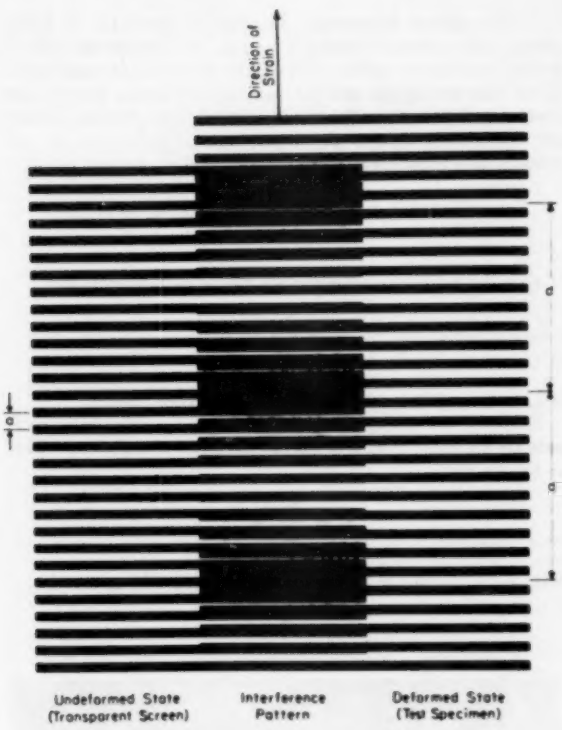


Fig. 1 Elongation perpendicular to the screen lines

It is clear that, in this case, the interference lines are parallel to the ruling and that the distance d must be measured perpendicularly to that direction.

It is not always possible to determine at once in every point whether the strain is elongation or compression. If, however, the sign of the deformation is known at one point, it is possible to determine the state of deformation in all neighboring points, because, in a homogeneous material, the deformation cannot show a discontinuous transition from elongation to compression. The

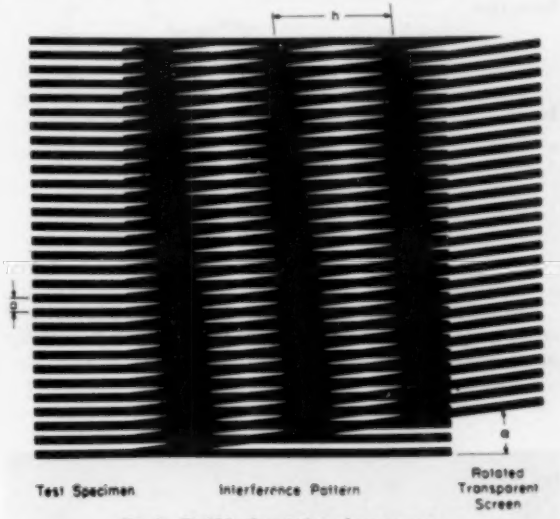
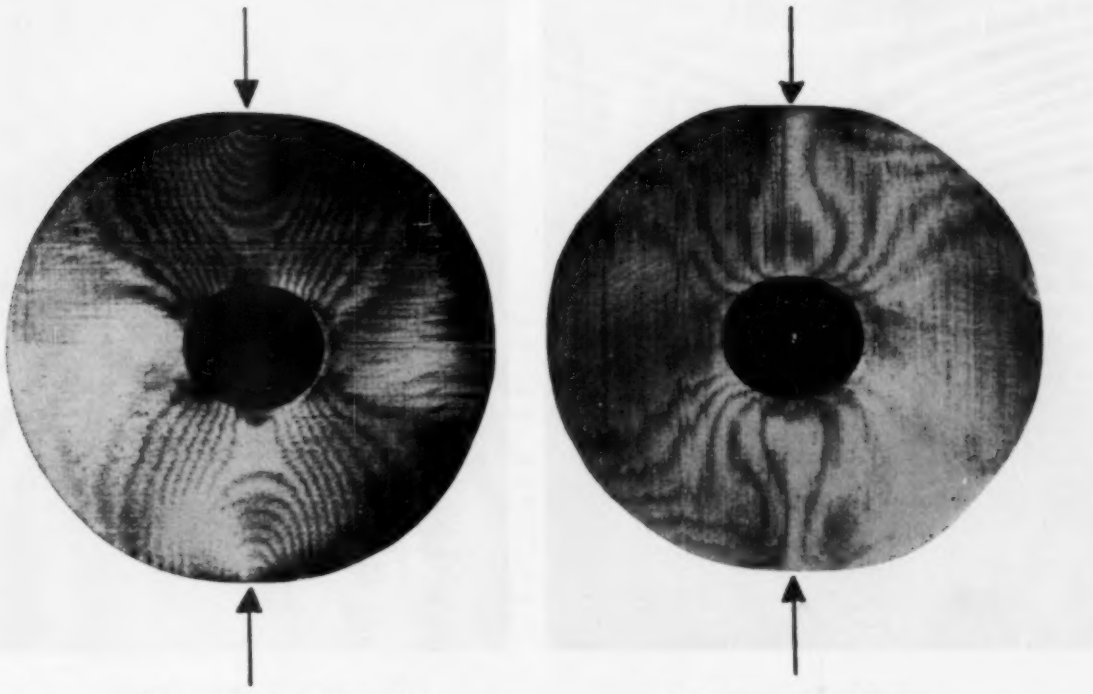


Fig. 2 Rigid-body rotation of one screen



a Lines perpendicular to load direction

b Lines parallel to load direction

Fig. 3 Compression test on ring

sign of the deformation in this point can now be found by measuring with a microscope the distance between two or more lines. However, this sign is usually known a priori in at least one point. For instance, from the nature of the problem, it is known at the point of application of load.

(b) Rigid body rotation of one screen. If the master screen is slightly rotated with regard to the undeformed screen on the specimen, interference lines appear, Fig. 2.

If we call h the distance between two neighboring interference lines, then

$$h = \frac{a}{\tan \alpha} + a \tan \frac{\alpha}{2}$$

hence

$$\sin \alpha = \frac{a}{h}$$

For small rotations this can be written

$$\alpha = \frac{a}{h}$$

The rotation-interference lines are slightly inclined (about $\alpha/2$) on the grid and the distance h must be measured parallel with the screen lines of the test specimen.

(c) Homogeneous deformation. In order to determine, in plane strain, both principal strains ϵ_1 and ϵ_2, two orthogonal sets of screen lines are to be provided on the surface of the specimen. If we take orthogonal axes OX and OY, the screen lines on the master grid being parallel to OX, then the interference fringes are the contour lines of the displacement function $v(x, y)$. It follows that, for small strains,

$$e_y = \frac{\partial v}{\partial y} = \frac{a}{d_y}$$

if d_y is the y-distance between two interference fringes, and

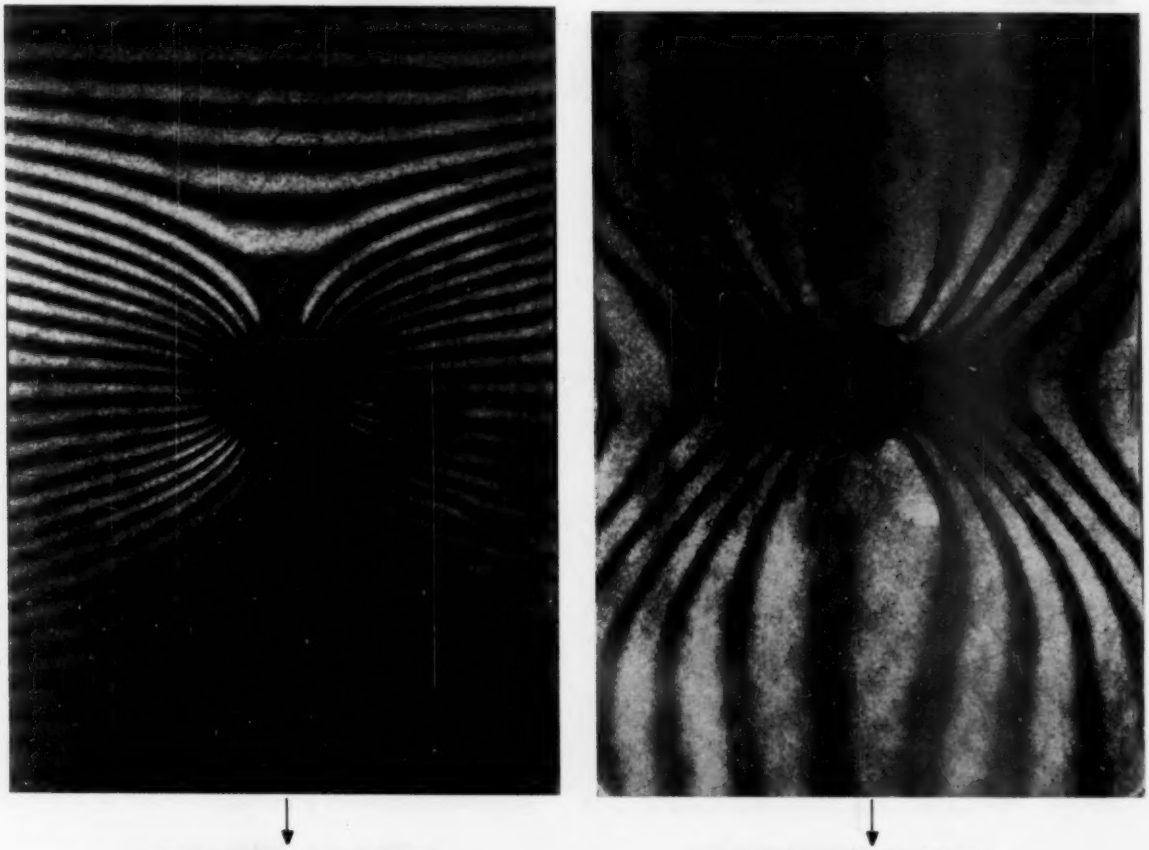
$$\alpha_z = \frac{\partial v}{\partial x} = \frac{a}{h_x}$$

if h_x is the x-distance between two interference fringes.

Similarly, the interference fringes of a grid parallel with the Y axis are the contour lines of the displacement $u(x, y)$, from which it follows that

$$e_x = \frac{\partial u}{\partial x} = \frac{a}{d_x}$$

if d_x is the x-distance between two interference fringes, and



a Lines perpendicular to load direction

b Lines parallel to load direction

Fig. 4 Tensile specimen with hole

$$\alpha_y = \frac{\partial u}{\partial y} = \frac{a}{h_y}$$

if h_y is the y-distance between two interference fringes.

We are thus able to determine the three strain components e_x, e_y, and $\gamma_{xy} = \alpha_x - \alpha_y$ for a homogeneous plane strain. The principal strains ϵ_1 and ϵ_3, and their direction θ are then given by:

$$\epsilon_1, \epsilon_3 = {}^1/_2\left[(e_x + e_y) \pm \sqrt{(e_x - e_y)^2 + \gamma_{xy}^2}\right]$$

and

$$\tan 2\theta = \frac{\gamma_{xy}}{e_x - e_y}$$

If the strain is not homogeneous, these formulas give the average values of the strain components in a small area, limited by the interference fringes.

A more rigorous mathematical treatment of the plane homogeneous deformation is given in the Appendix.

IV Results of Tests

A Compression Test on a Ring

A mild steel ring, 75.0 mm outside, 20.0 mm inside diam, 23.0 mm thick, and loaded in compression was the first specimen used to test the technique.

A set of equally spaced lines, 0.2 mm apart, were scribed on the polished surfaces with the help of a small dividing machine. One set of the lines was in the load direction, the other one perpendicular to it. The specimen was placed on a carriage and moved forward with a micrometer screw in steps of 0.2 mm. A cutting tool, fixed on a sledge, drew a very sharp and regular trace, ±0.05 mm wide, on the surface of the specimen. The spacing a between two lines, however, sometimes showed slight deviations, which resulted in an irregular Moiré pattern.

After a compressive load of 36,800 lb the interference fringes were obtained by superimposing a transparent screen with lines 0.1 mm apart on the test specimen, Fig. 3(a and b).

The picture shows that the deformation in the ring under the contact points is compression of high magnitude, accompanied by shear of two wedge-shaped volumes inward to the hole.

B Tensile Test on a Mild Steel Bar With a Central Circular Hole

The mild steel specimen used in this test, Fig. 4, was machined from a large hot-rolled ${}^1/_2$-in. plate. Its dimensions in inches are: width, 3.480; thickness, 0.470; radius of hole, 0.314. The 0.2 per cent yield strength, measured on a cylindrical test specimen, was 28,900 psi.

Before the test, both surfaces were polished and provided with a 133 lines-per-inch ruling photoengraved onto the surface. One side had the lines perpendicular to the axial load, the other side had the lines parallel to the load direction.

The bar was pulled in tension up to 58,200 lb (43,400 psi average on net section) and the corresponding interference pictures were taken, Fig. 4.

A square grid was drawn on the prints and the distances between the interference lines were measured in the x and y-directions along these lines. These data permitted the calculation of the average strain between two interference lines, shown by the stepped line graph, Fig. 5. These original curves, when averaged to one quadrant, yielded the curves of equal elongation e_x, Fig. 6, and equal compression e_y, Fig. 7. The strain concentration, along $x = 0$, is defined by:

$$\frac{e_x \text{ at the edge of the hole}}{\text{Average } e_x \text{ across minimum section}} = \frac{0.250}{0.066} = 3.8$$

This value may be much larger for smaller hole radii, and can be highly significant in the performance of strain-sensitive metals.

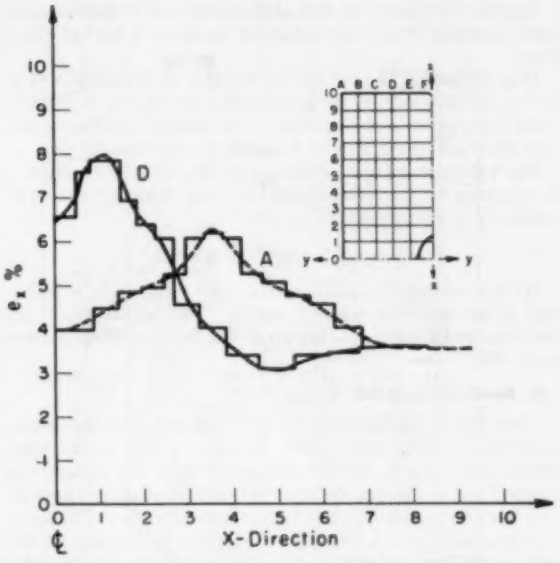


Fig. 5 Percentage elongation in x-direction along co-ordinate lines A and D

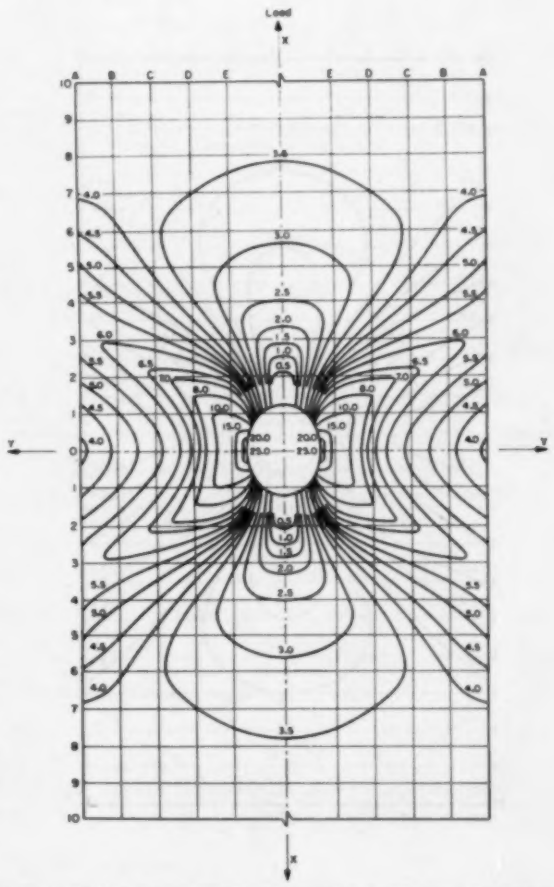


Fig. 6 Plastic deformation of tensile specimen with hole. Percentage elongation in x-direction.

Kawata [6] in his work with photoelastic coated specimens obtained a similar strain distribution for the case of a flat bar with a hole.

It is interesting to note the distribution of the maxima in e_x and e_y on each side of the longitudinal center line of the bar. They reflect clearly the existence of a cross-shaped depression, that often is found in largely deformed flat tensile bars.

Fig. 8 shows the variations of e_x, e_y, and e_z for $x = 0$. Here, e_z is measured by the reduction in thickness along $x = 0$. The constancy of volume is expressed by:

$$(1 + e_x)(1 + e_y)(1 + e_z) = 1$$

We can calculate from this equation the value e_z^* by substitution of the measured values e_x and e_y. The calculated e_z^* and the measured e_z-values are shown in Fig. 8. Both values agree fairly well.

C Tensile Specimen With Notches

A flat tensile specimen in mild steel had the following dimensions (in inches); width 2.000, thickness 0.052, notch radius 0.0625, notch depth 0.1875. Before the test, one surface was polished and a 133 lines-per-inch screen photoengraved onto it The lines were perpendicular to the load direction. The specimen was loaded until a crack appeared at the right notch; at this moment the load was 4800 lb or 51,900 psi average on the net surface. The interference picture of the deformation is shown in Fig. 9. The strain curves along vertical lines drawn on the print were derived from the spacings between the interference fringes along these lines. These curves permitted us to plot the lines of equal e_x elongation on the specimen, Fig. 10.

An interesting feature of this strain distribution is the low deformation region at the center of the specimen, which is bounded by areas of higher elongation.

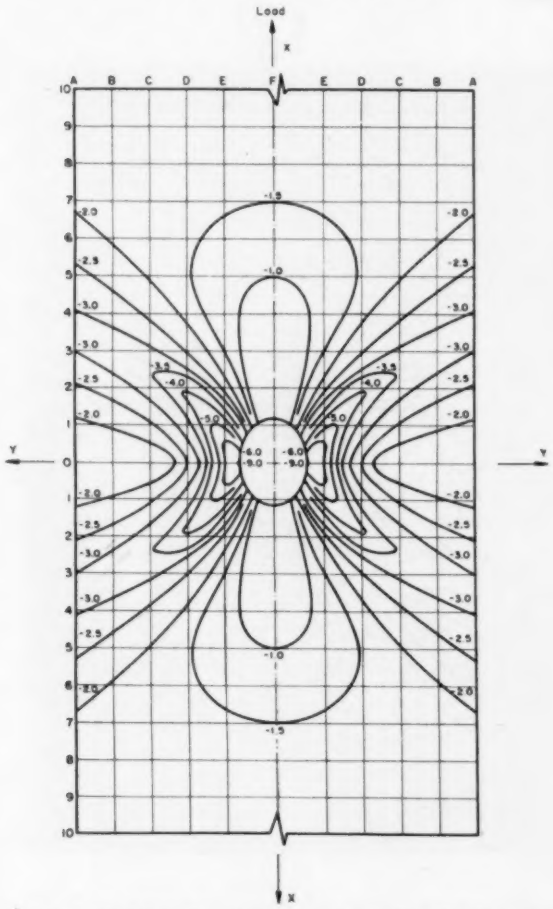


Fig. 7 Plastic deformation of tensile specimen with hole. Percentage elongation in y-direction.

References

1 B. M. Shepard and R. Weller, "Displacement Measurements by Mechanical Interferometry," *Proceedings, Society for Experimental Stress Analysis*, vol. VI, no. 1, 1948, pp. 35–38.

2 J. D. C. Crisp, "The Measurement of Plane Strains by a Photo Screen Method," *Proceedings, Society for Experimental Stress Analysis*, vol. XV, no. 1, 1957, pp. 65–76.

3 F. K. Ligtenberg, "The Moiré Method, a New Experimental Method for the Determination of Moments in Small Slab Models," *Proceedings, Society for Experimental Stress Analysis*, vol. XII, no. 2, 1955, pp. 83–98.

4 R. Thiry and P. Dantu, "Exposé Général d'une Méthode Expérimental de Détermination Directe des Déformations dans un Solide," Ninth International Congress of Applied Mechanics, Proceedings, vol. VIII, 1957, pp. 490–499.

5 J. A. Miller, "Improved Photogrid Techniques for Determination of Strain Over Short Gage Lengths," *Proceedings, Society for Experimental Stress Analysis*, vol. X, no. 1, 1952, pp. 29–34.

6 K. Kawata, "Analysis of Elasto-Plastic Behaviour of Metals by Means of Photoelastic Coating Method," *Journal, Scientific Research Institute*, no. 1473, vol. 52, 1958, pp. 17–40.

7 A. Nadai, "Theory of Flow and Fracture of Solids," McGraw-Hill Book Company, Inc., New York, N. Y., second edition, 1950, chapter 12.

APPENDIX

Strain Components e_x, e_y, γ

We have shown how we can visualize the plastic deformation in a body and how we can deduce the principal strains in the case of

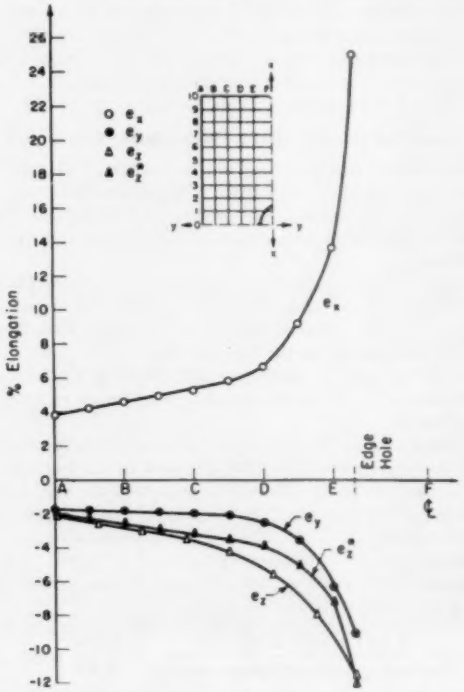


Fig. 8 Variations of e_x, e_y, e_z, and e_z^* along $x = 0$ in tensile specimen with hole

an infinitesimal deformation. In the case of a finite homogeneous plain strain, the second-order terms can no longer be neglected and other formulas are to be used.

We provide two orthogonal sets of screen lines on the surface of the body before deformation, one in the x-direction and one in the y-direction, Figs. 11 and 12. The Moiré pattern is obtained by using a fixed screen having a similar set of screen lines parallel with the x and y axes, which are fixed and do not follow the deformation.

From the Moiré fringes one measures the following quantities:

Fig. 11: d_y = y-distance between two neighboring fringes
h_x = x-distance between two neighboring fringes
β_x = angle between Ox and fringes
Fig. 12: d_x = x-distance between two neighboring fringes
h_y = y-distance between two neighboring fringes
β_y = angle between Oy and fringes

The real strain is:

$$\epsilon_x = \frac{OB' - OB}{OB}; \quad \epsilon_y = \frac{OA' - OA}{OA} \tag{1}$$

while we may call "apparent strain," the quantities:

$$e_x = \frac{OB'' - OB}{OB}; \quad e_y = \frac{OA'' - OA}{OA} \tag{2}$$

The apparent strain e_x and e_y are calculated from the measured quantities d_x and d_y.

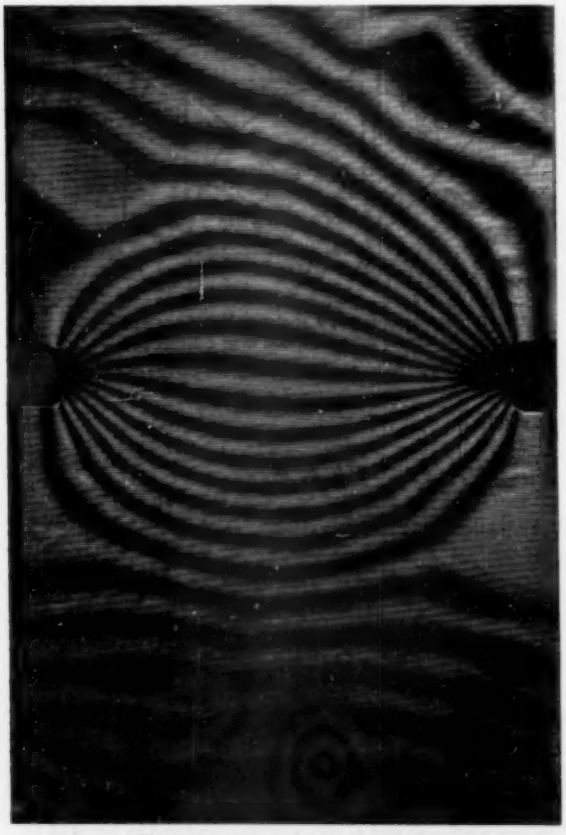


Fig. 9 Tensile specimen with notches

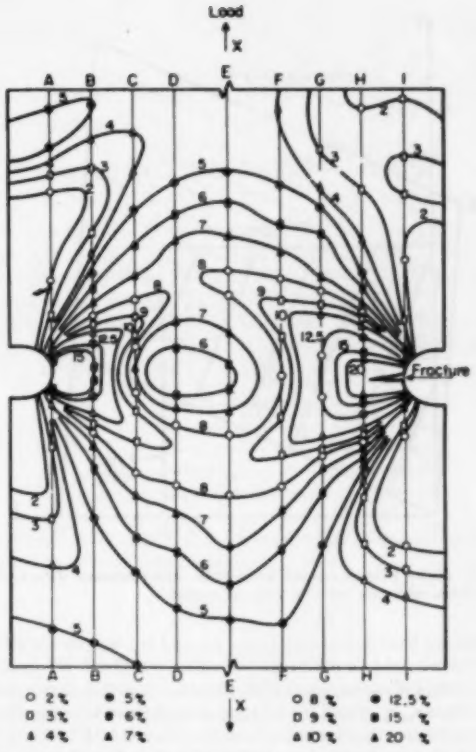


Fig. 10 Plastic deformation of notched tensile specimen at the moment of fracturing. Percentage elongation in x-direction.

In the case of extension

$$e_x = \frac{a}{d_x - a} \quad \text{and} \quad e_y = \frac{a}{d_y - a} \tag{3}$$

In the case of compression

$$e_x = \frac{-a}{d_x + a} \quad \text{and} \quad e_y = \frac{-a}{d_y + a} \tag{3a}$$

The *unit shear* [7] in point O, Fig. 13, is defined by:

$$\gamma = \tan \psi = \tan (\alpha_x - \alpha_y) \tag{4}$$

This can be calculated as follows:

Consider the triangle RST, Fig. 11. We obtain:

$$h_x = \frac{a}{\tan \alpha_x} - \frac{a}{\tan \beta_x}$$

hence

$$\tan \alpha_x = \frac{a}{h_x + \frac{a}{\tan \beta_x}} \tag{5}$$

In the case of extension we have:

$$h_x = \frac{-d_y}{\tan \beta_x} \tag{6}$$

From equations (3), (5), and (6) it follows that, in any case:

$$\tan \alpha_x = -e_y \tan \beta_x \tag{7}$$

Similarly, we obtain:

$$\tan \alpha_y = -e_x \tan \beta_y \tag{7a}$$

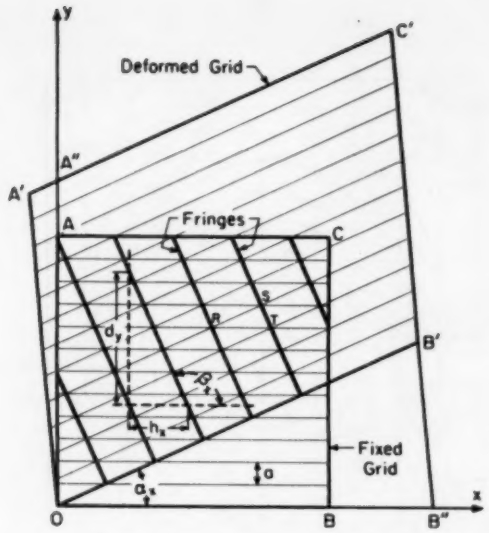


Fig. 11 Moiré fringes caused by a finite homogeneous deformation. Screen lines originally parallel with the x-axis.

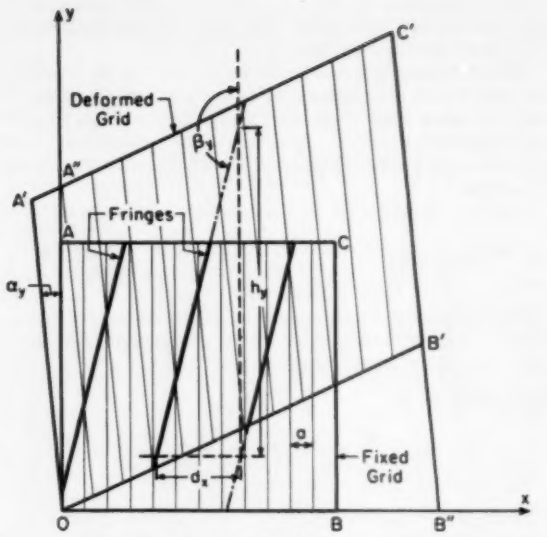


Fig. 12 Moiré fringes caused by a finite homogeneous deformation. Screen lines originally parallel with the y-axis.

When we have calculated e_x, e_y, α_x, and α_y, we possess all the elements we need to construct the rhomboid $OA'C'B'$ obtained by distortion of the square $OACB$, Fig. 13. It is therefore possible to calculate ϵ_x, ϵ_y, γ and the rotation ω from these four quantities.

Let us assume that the transformation from $OACB$ to $OA'C'B'$ is brought about by the linear vector function $\mathbf{B}$ so that the new co-ordinates x' are given

$$\mathbf{x}' = \mathbf{B}\mathbf{x} \tag{8}$$

or

$$\begin{Vmatrix} x' \\ y' \end{Vmatrix} = \begin{Vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{Vmatrix} \cdot \begin{Vmatrix} x \\ y \end{Vmatrix}$$

This means that

$$\begin{aligned} x' &= b_{11}x + b_{12}y \quad \text{and} \\ y' &= b_{21}x + b_{22}y \end{aligned} \tag{9}$$

The co-ordinates of point B'' are obtained by substituting $x = na$, $y' = 0$ in (9), $(OB = na)$.

$$OB'' = x' = \frac{b_{11}b_{22} - b_{21}b_{12}}{b_{22}}\, na = \frac{|\mathbf{B}|}{b_{22}}\, na \tag{10}$$

If we express OB'' as a function of the apparent strain, we find

$$1 + e_x = \frac{|\mathbf{B}|}{b_{22}}$$

and in the same way, we find (11)

$$1 + e_y = \frac{|\mathbf{B}|}{b_{11}}$$

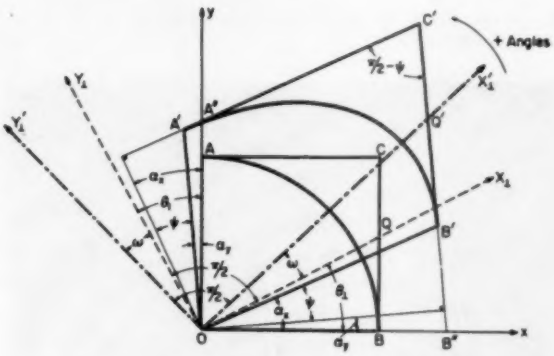
The co-ordinates of point B' are found by putting $y = 0$ in equation (9).

$$x' = b_{11}x$$

$$y' = b_{21}x$$

Then

$$\tan \alpha_x = \frac{b_{21}}{b_{11}}$$



Ox, Oy fixed orthogonal axes
OX_1, OY_1 directions of principal strains in the undeformed axes
OX_1', OY_1' directions of principal strains after deformation

Fig. 13 Finite homogeneous plane deformation

and similarly (12)

$$\tan \alpha_y = \frac{-b_{12}}{b_{22}}$$

The equations (11) and (12) yield

$$\left.\begin{aligned} b_{11} &= \frac{1 + e_x}{1 + \tan \alpha_x \tan \alpha_y} \\ b_{22} &= \frac{1 + e_y}{1 + \tan \alpha_x \tan \alpha_y} \\ b_{12} &= \frac{-\tan \alpha_y (1 + e_y)}{1 + \tan \alpha_x \tan \alpha_y} \\ b_{21} &= \frac{\tan \alpha_x (1 + e_x)}{1 + \tan \alpha_x \tan \alpha_y} \end{aligned}\right\} \tag{13}$$

The real strains ϵ_x and ϵ_y can now be calculated in terms of b_{11}, b_{12}, b_{21}, and b_{22}.

The co-ordinates of point B' are obtained by putting $x = na$, $y = 0$ in equation (9) ($OB = na$).

$$x' = b_{11}na$$

$$y' = b_{21}na$$

Hence:

$$OB' = \sqrt{b_{11}^2 + b_{21}^2}\, na$$

The real strain ϵ_x is obtained from

$$(1 + \epsilon_x)^2 = \left(\frac{OB'}{OB}\right)^2 = b_{11}^2 + b_{21}^2$$

Similarly (14)

$$(1 + \epsilon_y)^2 = b_{12}^2 + b_{22}^2$$

Substitution of equation (13) in (14) yields:

$$(1 + \epsilon_x)^2 = (1 + e_x)^2 \frac{1 + \tan^2 \alpha_x}{(1 + \tan \alpha_x \tan \alpha_y)^2} \quad \text{and}$$

$$(1 + \epsilon_y)^2 = (1 + e_y)^2 \frac{1 + \tan^2 \alpha_y}{(1 + \tan \alpha_x \tan \alpha_y)^2} \tag{15}$$

or:

$$1 + \epsilon_x = (1 + e_x) \frac{\cos \alpha_y}{\cos (\alpha_x - \alpha_y)} \quad \text{and}$$

$$1 + \epsilon_y = (1 + e_y) \frac{\cos \alpha_x}{\cos (\alpha_x - \alpha_y)} \tag{16}$$

If only one grid of parallel lines is scribed on the specimen, the Moiré pattern allows us to calculate, for instance, ϵ_x and α_y, but not ϵ_x.

The Moiré fringes can be brought nearer to each other, when we give the master screen a rotation α with respect to its original parallel position. A rigid body rotation: $\omega = \alpha, \epsilon_x = \epsilon_y = \gamma = 0$, results in an apparent strain e_x' deduced from (16):

$$1 + e_x' = \frac{1}{\cos \alpha} \tag{17}$$

If the body afterward undergoes an additional homogeneous plane deformation, then the values ϵ_x, ϵ_y, and γ, can still be calculated from (16) and (4). When only one set of parallel lines is drawn, the measured apparent strain e_x'' can be corrected with the additional e_x' by taking:

$$e_x = e_x'' - e_x' \tag{18}$$

Principal Strain-Components and Directions

The transformation expressed by the linear vector function $\mathbf{B}$ can be separated into a rigid-body rotation $\mathbf{R}$ and a pure strain $\mathbf{C}$:

$$\mathbf{B} = \mathbf{RC} \tag{19}$$

or:

$$\begin{Vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{Vmatrix} = \begin{Vmatrix} \cos \omega & -\sin \omega \\ \sin \omega & \cos \omega \end{Vmatrix} \cdot \begin{Vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{Vmatrix} \tag{20}$$

with $c_{21} = c_{12}$.

This requires:

$$\tan \omega = \frac{b_{21} - b_{12}}{b_{11} + b_{22}} \tag{21}$$

or from equation (13):

$$\tan \omega = \frac{(1 + e_x) \tan \alpha_x + (1 + e_y) \tan \alpha_y}{(1 + e_x) + (1 + e_y)}$$

We thus find the numerical value of the rotation.

The symmetric linear vector function $\mathbf{C}$, expressing a pure strain without rotation, allows us to calculate the principal strain direction θ_1 and the principal strains ϵ_1 and ϵ_2.

If we change from the fixed orthogonal axes xOy to the fixed orthogonal axes XOY, Fig. 13, the angle between Ox and OX being θ, the linear vector function $\mathbf{B}$ is replaced by $\mathbf{B}'$, given by:

$$\mathbf{B}' = \mathbf{UBU}^* \tag{22}$$

with:

$$\mathbf{U} = \begin{Vmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{Vmatrix}, \quad \mathbf{U}^* = \begin{Vmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{Vmatrix}$$

One verifies, with the aid of the first formula (21), that this leaves the rotation ω unchanged, and that

$$\mathbf{B}' = \mathbf{RC}'$$

where (23)

$$\mathbf{C}' = \mathbf{UCU}^*$$

$\mathbf{R}$ is the same as in (19) and $\mathbf{U}$ is the same as in (22). It is possible to choose $\mathbf{U}$ (or θ) such that, in the new axes XOY:

$$\gamma' = 0$$

If this is the case, the corresponding directions OX and OY are called *principal directions*. For the components of the symmetric linear vector function $\mathbf{C}$, (4) and (13) require that:

$$c'_{12} = c'_{21} = 0$$

Substituting these values into the second equation (23) one finds:

$$\tan 2\theta_1 = \frac{2c_{21}}{c_{22} - c_{11}} \tag{24}$$

Solving (19) for the components of $\mathbf{C}$, and substituting this in (24):

$$\tan 2\theta_1 = 2 \frac{b_{12}b_{11} + b_{22}b_{21}}{-b_{11}^2 + b_{21}^2 + b_{12}^2 - b_{22}^2}$$

and finally, from (13):

$$\tan 2\theta_1 = 2 \frac{(1 + e_x)(1 + e_y)(\tan \alpha_y - \tan \alpha_x)}{(1 + e_x)^2(1 + \tan^2 \alpha_x) - (1 + e_y)^2(1 + \tan^2 \alpha_y)} \tag{25}$$

The direction corresponding to $0 < \theta_1 < \pi/2$ is the first principal direction OX_1, Fig. 13.

The principal strain ϵ_1 can be deduced from (14):

$$1 + \epsilon_1 = c_{11}'$$

by calculating c_{11}' from (19) and (23). The result is more readily obtained by considering the triangle OQB, transformed into $OQ'B'$, Fig. 13, and writing:

$$\overline{OQ}'^2 = \overline{OB}'^2 + \overline{B'Q}'^2 - 2\overline{OB}' \cdot \overline{B'Q}' \cdot \cos \widehat{OB'Q'} \tag{26}$$

Observing that $\overline{OQ}^2 = \dfrac{\overline{OB}^2}{\cos^2 \theta_1} = \dfrac{\overline{BQ}^2}{\sin^2 \theta_1}$

and dividing both sides of (26) by this quantity, one obtains:

$$(1 + \epsilon_1)^2 = (1 + \epsilon_x)^2 \cos^2 \theta_1 + (1 + \epsilon_y)^2 \sin^2 \theta_1 + 2(1 + \epsilon_x)(1 + \epsilon_y) \cos \theta_1 \sin \theta_1 \sin (\alpha_x - \alpha_y) \quad (27)$$

The principal strain ϵ_2 in the direction OY_1, is found to be:

$$(1 + \epsilon_2)^2 = (1 + \epsilon_y)^2 \cos^2 \theta_1 + (1 + \epsilon_x)^2 \sin^2 \theta_1 - 2(1 + \epsilon_y)(1 + \epsilon_x) \cos \theta_1 \sin \theta_1 \sin (\alpha_x - \alpha_y) \quad (28)$$

One readily verifies that:

$$(1 + \epsilon_1)^2 + (1 + \epsilon_2)^2 = (1 + \epsilon_x)^2 + (1 + \epsilon_y)^2$$

which replaces $\epsilon_1 + \epsilon_2 = \epsilon_x + \epsilon_y$ which is valid for infinitesimal strain.

A Generalization of Cumulative Damage[1]

R. M. MAINS

Consulting Engineer, Knolls Atomic Power Laboratory, Schenectady, N. Y.[2]

In the application of various damage accumulation hypotheses to shock and vibration problems, a pattern of results was observed. A plot of allowable response (for the largest of a sequence of excitations) versus the number of excitations in the sequence produced straight lines on log-log paper for each value of damping. It seemed likely that a generalization could be found which would relate the allowable response to the largest excitations with a load-distribution factor, a material-and-structure factor, and the material constants from fatigue tests. One form of this generalization is presented, together with some numerical examples of its use, and a discussion of its possible significance.

Introduction

In the course of several applications of damage accumulation hypotheses to shock and vibration problems [1],[3] there seemed to be a consistent pattern for the results. In each case the results could be presented as a graph of the allowable maximum stress (or plastic strain) response versus the applied number of shocks (or vibration sequences) for various values of structural damping. Changes in the distribution of load amplitudes shifted these curves as a group, also. As a consequence of these observations, it seemed possible that a generalization of damage accumulation could be made somewhat as follows:

Total damage accumulated = (amplitude of maximum response factor) × (distribution of load amplitudes factor) × (material and structure factor) (1)

If such a generalization could be made, then the writing of specifications and the design processes for shock and vibration could be greatly simplified and, at the same time, made more rational. It might even be possible to include factors for the damage contributed by other loads, such as thermal expansion, pressure cycles, mechanical loads, and the like. The following development leads to one form of such a generalization.

Fatigue Test Data

In order to begin, let us consider presently available relationships which could be put together to form the various factors in equation (1). First, it is reasonably well established [2, 3] that fatigue tests of steels give the relationship:

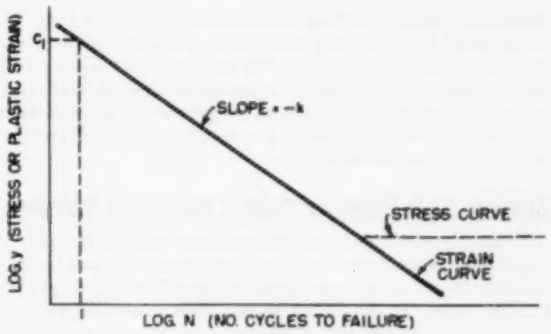


Fig. 1 Fatigue diagram

$$yN^k = C_1 \tag{2}$$

as shown in Fig. 1, in which

y = stress for stress fatigue or plastic strain for strain fatigue
N = number of cycles to failure at y
k = $1/_8$ for stress fatigue, or $1/_2$ for plastic strain fatigue
C_1 = constant for each material

Structural Response to Transient Loads

Second, it is well established [4] that elastic structural responses to transient loading of a structure have the form:

$$y = e^{-\xi\omega_0 t} C_2 \sin \omega_1 t + y_0(t) \tag{3}$$

as shown in Fig. 2, in which

y = stress or strain response
ξ = critical damping ratio
ω_1 = natural frequency of the responding mode
C_2 = initial amplitude, a function of the load
ω_0 = $\omega_1/\sqrt{1-\xi^2}$
$y_0(t)$ = stress or strain response to steady load or slowly (compared to $2\pi\omega_1$) varying load

If y in equation (3) is not large enough to involve yielding of the material, then y, C_2, and y_0 may represent either stresses or strains. If y does involve yielding, then Equation (3) is not exact because the rate of amplitude decay is likely to be faster than $e^{-\xi\omega_0 t}$ would provide. However, if the amplitude decay curve for y can be measured or calculated, then a value of $\xi\omega_0$ can be found which will satisfactorily approximate the decay curve. (A good rule of thumb is to take ξ as the reciprocal of 2 times the number of cycles required to decay to 5 per cent of the initial amplitude.) So for the purpose of this generalization, we will assume that y in equation (3) can also represent plastic strain responses.

---

[1] Legal notice: This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

(*a*) Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

(*b*) Assumes any liabilities with respect to the use of, or for damage resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

[2] Operated for the United States Atomic Energy Commission by the General Electric Company under Contract No. W-31-109 Eng-52.

[3] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 26, 1959. Paper No. 59—Met-1.

Damage Accumulation

Third is the matter of damage accumulation, which is not well established at all. The assumption of linear damage accumulation (Palmgren 1924, Serensen 1944, Miner 1945) usually called Miner's hypothesis [5] has the form:

$$\delta = \sum_i \frac{n_i}{N_i} \tag{4}$$

in which

N_i = the number of cycles to failure at a particular stress or strain for the virgin material

n_i = the number of cycles applied at that particular stress or strain (equal to one for individual cycles)

δ = the proportion of damage accumulated in a sequence of i groups of stresses or strains.

Miner's hypothesis was that failure occurred at $\delta = 1$, but more recent work [6, 7, 8] shows that failure occurs for values of δ as small as 0.1 and as large as 10. The correct value to use for δ depends upon a number of factors which are still being investigated, so we shall proceed with deriving the *form* of the generalization and determine the importance of a specific value for δ when we see how it affects the results.

Stress or Strain Range per Cycle of Response to Transients

To put the available elements from equations (2), (3), and (4) together, consider the range of amplitude covered by each cycle of equation (3). In Fig. 2, this range is the difference between the positive peaks, (1), (3), and so on, and their succeeding negative peaks, (2), (4), and so on. Mathematically, this difference is the difference between y_r at $\omega_1 t = \frac{\pi}{2}(4r-1)$ and

$$y_{r-1} \quad \text{at} \quad \omega_1 t = \frac{\pi}{2}(4r-3).$$

This is

$$\left.\begin{aligned} y_R = y_{r-1} - y_r = {} & e^{-\xi\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}} C_2 \sin\frac{\pi}{2}(4r-3) \\ & - e^{-\xi\frac{\pi}{2}\frac{4r-1}{\sqrt{1-\xi^2}}} C_2 \sin\frac{\pi}{2}(4r-1) \end{aligned}\right\} r = 1, 2, 3 \ldots$$

$$y_R = C_2\left[\exp\left(-\xi\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}\right) + \exp\left(-\xi\frac{\pi}{2}\frac{4r-1}{\sqrt{1-\xi^2}}\right)\right]_{r=1,2,3\ldots} \tag{5}$$

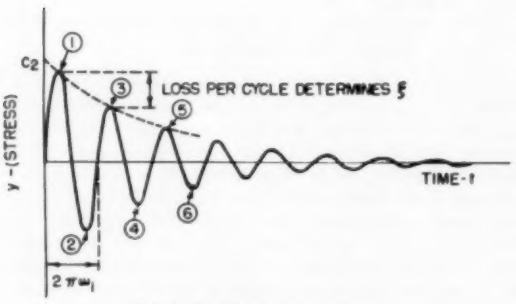


Fig. 2 Transient response curve

$$y_R = C_2\left[\left(1 + e^{-\frac{\pi\xi}{\sqrt{1-\xi^2}}}\right)\exp\left(-\xi\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}\right)\right]_{r=1,2,3\ldots} \tag{6}$$

Note that r is the number of the cycle of stress or strain in the particular series of stresses or strains resulting from the application of one transient load.

In order to determine how many cycles in the series are of significance, consider the ratio of y_R for $r = 1$ and for $r = 1/\xi$

$$\frac{y_{1/\xi}}{y_1} = \frac{\exp\left(-\xi\frac{\pi}{2}\frac{4-3\xi}{\xi\sqrt{1-\xi^2}}\right)}{\exp\left(-\xi\frac{\pi}{2}\frac{1}{\sqrt{1-\xi^2}}\right)} = \exp\left(\frac{-2\pi(1-\xi)}{\sqrt{1-\xi^2}}\right) \tag{7}$$

Then for various values of ξ,

ξ	1/2	1/4	1/8	1/16	1/32	1/64
$y_{1/\xi}/y_1$	0.0268	0.00470	0.00342	0.00263	0.00227	0.00207

and it seems reasonable to say that carrying r out to $1/\xi$ will cover the significant stresses or strains in the damage accumulation.

Damage Accumulation per Transient Load

Now let us hold equation (6) in mind while we look again at equation (2), which can be rewritten as

$$1/N = (y/C_1)^{1/k} \tag{8}$$

If we consider further the portion of damage contributed by a single cycle in equation (4),

$$\delta_i = 1/N_i \tag{9}$$

and substitute equation (8) into equation (9), we get:

$$\delta_i = \left(\frac{y}{C_1}\right)^{\frac{1}{k}} \tag{10}$$

Now substitute equation (6) into equation (10) and obtain:

$$\delta_i = \left(\frac{C_2}{C_1}\right)^{\frac{1}{k}}\left[\left(1 + e^{\frac{-\pi\xi}{\sqrt{1-\xi^2}}}\right)\exp\left(-\xi\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}\right)\right]^{\frac{1}{k}} \tag{11}$$

as the portion of damage contributed in each individual cycle r of the damped train of response to a transient. If we agree to ignore the cycles after $r = 1/\xi$, then the sum of the damage contributed by the train of cycles of response to a single transient becomes:

$$\delta_s = \sum_{r=1,2,3\ldots 1/\xi}\left(\frac{C_2}{C_1}\right)^{\frac{1}{k}}\left(1 + e^{\frac{-\pi\xi}{\sqrt{1-\xi^2}}}\right)^{\frac{1}{k}}\exp\left(-\frac{\xi}{k}\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}\right) \tag{12}$$

This can be rewritten as:

$$\delta_s = \left(\frac{C_2}{C_1}\right)^{\frac{1}{k}}\left[\left(1 + e^{\frac{-\pi\xi}{\sqrt{1-\xi^2}}}\right)^{\frac{1}{k}} \sum_{r=1,2,3\ldots 1/\xi}\exp\left(-\frac{\xi}{k}\frac{\pi}{2}\frac{4r-3}{\sqrt{1-\xi^2}}\right)\right] \tag{13}$$

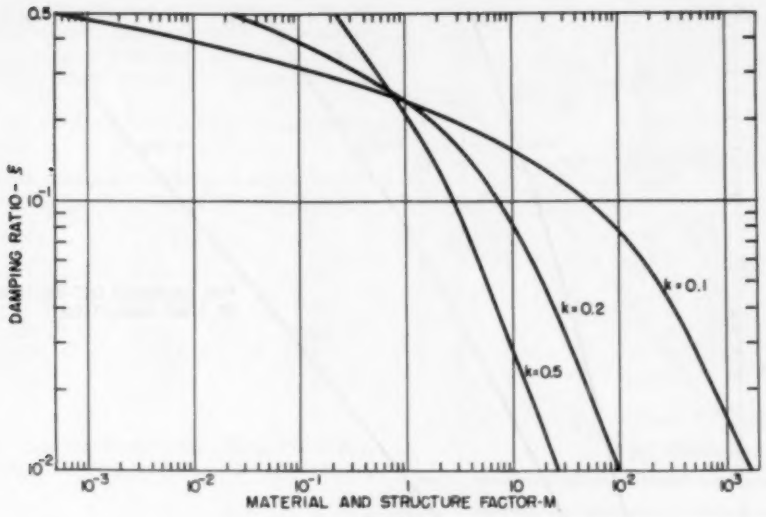


Fig. 3

The term in brackets in equation (13) depends only upon ξ, k, and how many cycles r are counted, so that it is clearly a factor dependent upon the material and the structure. Let us then define a material and structure factor M as:

$$M = \left(1 + e^{\frac{-\pi\xi}{\sqrt{1-\xi^2}}}\right)^{\frac{1}{k}} \sum_{r=1,2,3\ldots\frac{1}{k}} \exp\left(-\frac{\xi}{k}\,\frac{\pi}{2}\,\frac{4r-3}{\sqrt{1-\xi^2}}\right) \quad (14)$$

and consider it for use in equation (1). Fig. 3 shows a plot of M as a function of the critical damping ratio ξ and various values of k.

Damage Accumulation From a Series of Transients

Next, let us rewrite equation (13) as:

$$\delta_s = M\left(\frac{C_2}{C_1}\right)^{\frac{1}{k}} \quad (15)$$

and recall that δ_s is the portion of damage accumulated by the *train* of cycles of response to one transient. To determine the effect of a series of transients, we can now let C_2 take on a series of values, each of which is the initial response to a different transient. Then the sum of the damage accumulated in m such transients is

$$\delta_T = M \times \sum_m \left(\frac{C_{2m}}{C_1}\right)^{\frac{1}{k}} \quad (16)$$

in which C_{2m} is the initial response for the mth transient. This equation can be rewritten as

$$\delta_T = M \times \sum_m \left(\frac{C_{2m}}{C_1} \cdot \frac{C_{\max}}{C_{\max}}\right)^{\frac{1}{k}} \quad (17)$$

in which $C_{\max}$ is the largest initial response in the series of transients. Since $C_{\max}/C_1$ is a constant factor, it can be brought outside the summation to give:

$$\delta_T = \left(\frac{C_{\max}}{C_1}\right)^{\frac{1}{k}} \times M \times \sum_m \left(\frac{C_{2m}}{C_{\max}}\right)^{\frac{1}{k}} \quad (18)$$

The part left within the summation is dimensionless, and depends only upon the *ratios* of the initial responses to the maximum initial response, the number of transients, and the constant k from equation (2). Consequently, we can now think of the summation term in equation (18) as a load distribution factor D and define it as follows:

$$D = \sum_m \left(\frac{C_{2m}}{C_{\max}}\right)^{\frac{1}{k}} \quad (19)$$

and consider it for use in equation (1).

Table 1 Load amplitude distributions for various total numbers of loads

(Normal probability or Gaussian distributions)

Intensity	Number of loads in intensity interval for total number n				
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
0	44.3	122.1	258.5	545.8	1453.0
10	40.9	110.1	232.2	482.6	1264.0
20	35.0	91.1	187.0	380.6	961.5
30	27.6	68.6	135.4	263.4	638.0
40	20.2	47.0	87.8	164.8	369.0
50	13.6	29.3	51.4	90.0	186.5
60	8.5	16.6	26.8	43.4	81.5
70	5.0	8.6	12.6	18.8	31.8
80	2.6	3.9	5.3	7.2	10.9
90	1.3	1.7	2.0	2.4	2.8
100	1.0	1.0	1.0	1.0	1.0

Effect of Factor *D*

In order to provide some feeling for the effect of factor D, its value was computed for a series of normal probability (Gaussian) distributions of various total numbers of transient loads. The amplitude distributions for the various series are given in Table 1, and a plot of D versus total number of these transient loads for three values of k is shown in Fig. 4. (In addition, it is worth while to observe that, for a steady sinusoidal load, D becomes simply m while the material factor M becomes unity. For a load applied only once slowly and removed slowly, both D and M become unity.)

Amplitude Factor

Equation (18) can now be rewritten as:

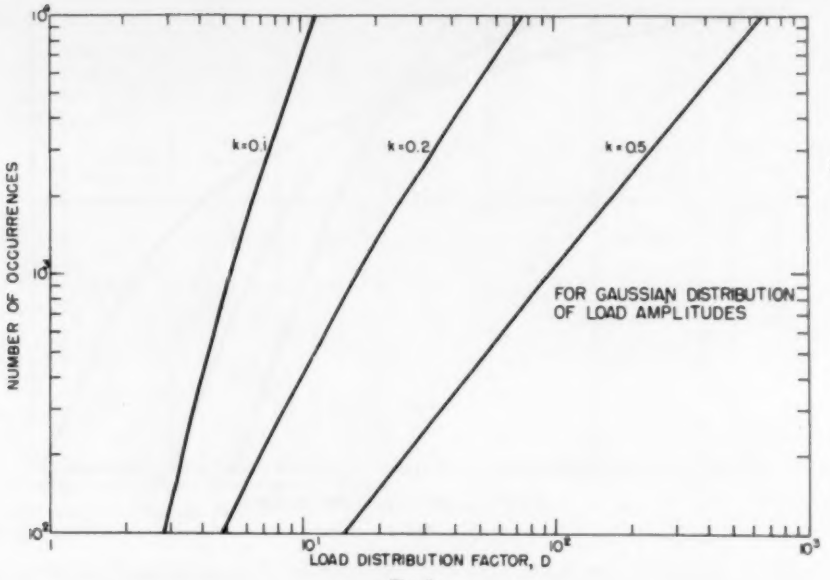


Fig. 4

$$\delta_T = \left(\frac{C_{max}}{C_1}\right)^{\frac{1}{k}} \times D \times M \tag{20}$$

In this equation, C_{max} is the maximum initial response, and C_1 and k are material constants from equation (2). This term is again dimensionless and qualifies as an amplitude of maximum response factor A defined as follows:

$$A = \left(\frac{C_{max}}{C_1}\right)^{\frac{1}{k}} \tag{21}$$

This completes our search for terms to put in equation (1), which now reads:

$$\delta_T = A \times D \times F \tag{22}$$

Recapitulation

To recapitulate before exploring further, let us examine equation (22):

1 The material-and-structure factor M is a constant, linear multiplier once the exponent k in equation (2), the damping ξ in equation (3), and the number of cycles to be counted in each transient response r have been determined.

2 The load-distribution factor D is a constant, linear multiplier once the number of separate transient disturbances m their shape distribution $\frac{C_{2m}}{C_{max}}$, and the exponent k in equation (2) are determined.

3 The amplitude-of-maximum-response factor A is dependent only upon the largest initial response C_{max} and the constants C_1 and k are established in equation (1).

4 The damage accumulation number δ_T represents the same thing as δ in equation (4), and should indicate failure for some value, presumably near 1.

Inclusion of Other Kinds of Loads

Equation (22) is then an answer to our quest as set forth at the beginning, and as defined verbally in equation (1). Suppose we are concerned with several different kinds of excitation each acting at separate times, such as shock, vibration, thermal cycles, pressure cycles, and the like. Then equation (22) would become:

$$\delta_{tot} = \sum_i (A_i \times D_i \times M_i) \tag{23}$$

in which the subscripts i denote the specific values of the various factors appropriate to each of the different kinds of excitation. We can go further by normalizing equation (23) and divide each A_i by the maximum value A_{max} which gives:

$$\delta_{tot} = A_{max} \sum_i \left(\frac{A_i}{A_{max}} \times D_i \times M_i\right) \tag{24}$$

Now since

$$A_{max} = \left(\frac{C_{max}}{C_1}\right)^{\frac{1}{k}}$$

from equation (21), equation (24) can be rearranged and solved for C_{max} to give:

$$C_{max} = C_1 \delta_{tot}^{\,k} \left[\sum_i \frac{A_i}{A_{max}} \times D_i \times M_i\right]^{-k} \tag{25}$$

If we establish a limiting value for δ_{tot} in equation (25), the equation then expresses the allowable maximum initial response $\bar{C}_{max}$ thus:

$$\bar{C}_{max} = C_1 \delta_{lim}^{\,k} \left[\sum_i \frac{A_i}{A_{max}} \times D_i \times M_i\right]^{-k} \tag{26}$$

which could be used as a design limit.

Precautions and Significance of Factors

In using equation (23) as an evaluation tool or equation (26) as a design tool, it should be noted that, if various kinds of loads are applied simultaneously, then their combination into cycles of various sorts must be done before the equations are applied. Such a procedure does not alter the validity of the generalization,

and it still provides a simplification over any other way known to the author to include the same considerations in an analysis. The validity of the generalization as such, however, must be checked against available service or test data before it can be accepted as correct.

The significance of the various factors in equation (26) would seem to be as follows:

1 The allowable total damage accumulation is a factor raised to the kth power. The significance of δ_{lim} in affecting $\bar{C}_{max}$ may be judged by:

	$\delta_{lim}{}^k$ for		
δ_{lim}	$k = 0.1$	0.2	0.5
0.3	0.887	0.717	0.548
0.5	0.927	0.859	0.707
0.8	0.978	0.956	0.895
1.0	1.000	1.000	1.000

in which we see that $\bar{C}_{max}$ is not particularly sensitive to δ_{lim}. This may help to explain why large variations in δ_{lim} are found experimentally.

2 Since the linear combinations of A_i, D_i, and M_i are summed and raised to the $-k$ power, and since k is less than 1.0, these factors are magnified in importance. Hence their proper determination is of considerable significance in establishing $\bar{C}_{max}$.

3 The material constants C_1 and k are of primary importance in establishing $\bar{C}_{max}$, as they should be.

4 Any term in equation (26) is eligible for the application of a safety factor, depending upon which uncertainties are to be so covered, and the resulting effect on $\bar{C}_{max}$ can be evaluated from the equation. Safety factors applied to A_i, D_i, or M_i are magnified in effect, while for δ_{lim} the effect is diminished. Where such safety factors should be used is a matter for the individual problem.

Effect of Mean Stress

In reviewing this generalization, Dr. W. E. Cooper[4] pointed out that the effect of mean stress (or strain) on fatigue life had not been included, and suggested that it could be accounted for as follows:

Mean stress (or strain) affects the fatigue life at a particular alternating stress (or strain), the same as if some different alternating stress (or strain) alone had been applied. That is to say that a combination of alternating stress (or strain) and mean stress (or strain) can be replaced by an equivalent alternating stress (or strain) which has the same damage effect as the combination. In equation (3), let C_2 be replaced by $C_2{}^*$, the equivalent alternating stress (or strain) with corresponding changes in C_{2m} in equation (16) and C_{max} in equation (17). Then

$$C_2{}^* = C_2[1 - f(y_{mean}, y_s)]^{-1} \tag{27}$$

in which

C_2 = initial amplitude in equation (3)
y_{mean} = mean stress (or strain) parameter
y_s = material parameter
$f(y_{mean}, y_s)$ = some function which produces the proper relationship between C_2 and $C_2{}^*$
$C_2{}^*$ = equivalent alternating stress (or strain)

When the actual combined mean and alternating stresses stay within the elastic range, and the modified Goodman diagram [10] is used for defining the effect of mean stress on fatigue life, then equation (27) becomes (see Fig. 5):

$$C_2{}^* = C_2\left[1 - \frac{y_{mean}}{y_s}\right]^{-1} \tag{28}$$

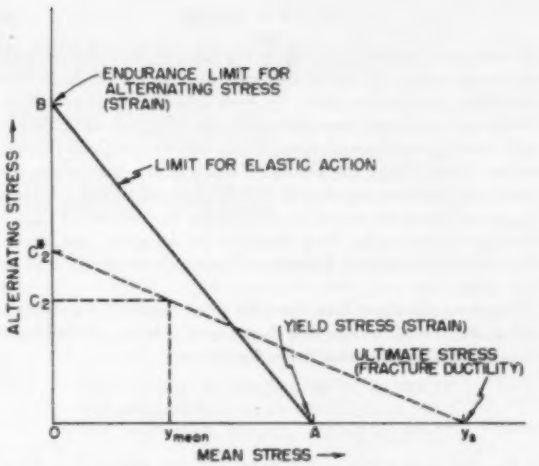


Fig. 5 Modified Goodman diagram for equivalent alternating stress

Notes: Ordinates and abscissas must have same units.
Area OAB can be stress or strain.
Area outside OAB must be strain or "fictitious elastic stress."

in which

y_{mean} = actual mean stress
C_2 = actual alternating stress amplitude
y_s = ultimate stress for the material

When strains are used instead of stresses, equation (28) still holds, with appropriate interchanges of strain for stress in the definitions. In this case, the strains need not be restricted to the elastic range, but may include plastic strains as well.

For cases in which the combined mean and alternating strains do not exceed 3 or 4 times the yield strain, then equation (28) can be used with the various parameters, particularly the mean stress parameter, defined as "fictitious elastic stresses" in the manner proposed by Langer [11].

In any of these cases, and whatever relationship is used to define the effect of mean stress on fatigue life, the results in terms of this generalization are:

1 C_{2m} and C_{max} in the load distribution factor, equation (19), are replaced by $C_{2m}{}^*$ and $C_{max}{}^*$, the equivalent alternating stresses (or strains).

2 C_{max} in the amplitude of maximum response factor, equation (21), is replaced by $C_{max}{}^*$.

3 $\bar{C}_{max}$ the allowable maximum initial response, equation (26), becomes $\bar{C}_{max}{}^*$, the equivalent alternating response value.

Numerical Illustrations

To provide some numerical illustrations of this generalization, the following forms of equation (2) were taken from the literature:

For carbon steel pipe [2]

$$\sigma\, N^{1/4} = 245{,}000 \tag{28}$$

For stainless steel pipe [3]

$$\epsilon_p\, N^{1/4} = 0.35 \tag{29}$$

For butt welds in structural steel plate [9]

[4] Consulting Engineer, Knolls Atomic Power Laboratory, General Electric Company.

$$\sigma\, N^{1/18} = 150{,}000 \qquad (30)$$

It was then assumed that it was desired to find the allowable maximum stress or strain response to a series of shock loads according to equation (26). In each case a normal probability distribution of shock load intensities was assumed, as in Table 1, with varying numbers of occurrences, and with various damping ratios. In addition, the allowable value of δ_{11m} was taken arbitrarily as 1.0 in keeping with Miner's hypothesis. Other values for δ_{11m} would have the effect of multiplying the results for $\bar{C}_{max}$ by the appropriate factor from the table in the section on Precautions and Significance of Factors. The results are given in Tables 2, 3, and 4.

These results show that, for 1000 total shocks and a structure and material with an effective damping of $\xi = 1/16$, the allowable response to the maximum shock would be:

82,300 psi for the material of equation (28)
0.024 strain for the material of equation (29)
77,000 psi for the material of equation (30)

If δ_{11m} had been taken as 0.8 instead of 1.0, these values would have been reduced by 4.4, 10.5, and 2.2 per cent, respectively. With either value of δ_{11m}, the allowable response to maximum shock would be much greater than the usual values associated with yield stress limits.

Table 2 Results of applying equation (26) for a Gaussian distribution of shock loads to $\sigma \times N^{1/18} = 245{,}000$, $\delta_{11m} = 1.0$

No. of shock loads m	Dist. factor D	Damp. ratio ξ	Material factor M	$\bar{C}_{max}$
200	6.82	1/2	0.0229	355,000 psi[a]
		1/4	0.825	174,000[a]
		1/8	4.98	121,000[a]
		1/16	14.25	98,000
		1/32	31.3	84,000
		1/64	61.7	73,000
500	11.14	1/2	0.0229	322,000[a]
		1/4	0.825	158,000[a]
		1/8	4.98	110,000
		1/16	14.25	89,000
		1/32	31.3	76,000
		1/64	61.7	66,000
1000	16.4	1/2	0.0229	298,000[a]
		1/4	0.825	146,000[a]
		1/8	4.98	102,000
		1/16	14.25	82,000
		1/32	31.3	67,000
		1/64	61.7	61,000
2000	25.2	1/2	0.0229	273,000[a]
		1/4	0.825	134,000[a]
		1/8	4.98	93,000
		1/16	14.25	75,000
		1/32	31.3	65,000
		1/64	61.7	56,000
5000	46.9	1/2	0.0229	242,000[a]
		1/4	0.825	98,000
		1/8	4.98	82,000
		1/16	14.25	67,000
		1/32	31.3	57,000
		1/64	61.7	50,000

[a] Should be treated as strain fatigue problem.

Table 3 Results of applying equation (26) for a Gaussian distribution of shock loads to $\epsilon_p \times N^{1/3} = 0.35$, $\delta_{11m} = 1.0$

No. of shock loads m	Dist. factor D	Damp. ratio ξ	Material factor M	$\bar{C}_{max}$
200	25.4	1/2	0.220	0.148 strain
		1/4	0.734	0.081
		1/8	2.27	0.046
		1/16	4.53	0.033
1000	92.6	1/2	0.220	0.078
		1/4	0.734	0.042
		1/8	2.27	0.024
		1/16	4.53	0.017
5000	365	1/2	0.220	0.039
		1/4	0.734	0.021
		1/8	2.27	0.012
		1/16	4.53	0.009

Table 4 Results of applying equation (26) for a Gaussian distribution of shock loads to $\sigma \times N^{1/18} = 150{,}000$, $\delta_{11m} = 1.0$

200	3.36	1/8	23.7	97,000 psi
		1/16	154	80,000
		1/32	447	72,000
		1/64	996	67,000
1000	5.10	1/8	23.7	93,000
		1/16	154	77,000
		1/32	447	69,000
		1/64	996	64,000
5000	9.00	1/8	23.7	88,000
		1/16	154	73,000
		1/32	447	65,000
		1/64	996	60,000

References

1 R. M. Mains, "An Application of Accumulative Damage Criteria," 25th Shock and Vibration Bulletin, Part II, 1957.

2 A. R. C. Markl, "Fatigue Tests of Piping Components," TRANS. ASME, vol. 74, 1952, pp. 287–303.

3 E. E. Baldwin, G. J. Sokol, L. F. Coffin, "Cyclic Strain Fatigue Studies, on AISI Type 347 Stainless Steel," *Proceedings, ASTM*, vol. 57, 1957, pp. 567–586.

4 S. P. Timoshenko, "Vibration Problems in Engineering," Van Nostrand and Company, New York, N. Y., 1937, p. 32 ff.

5 M. A. Miner, "Cumulative Damage in Fatigue," *Journal of Applied Mechanics*, vol. 21, TRANS. ASME, vol. 67, 1947, p. 159.

6 H. T. Corten and T. J. Dolan, "Cumulative Fatigue Damage," International Conference on Fatigue of Metals, session 3, paper 12, I. Mech. E. and ASME, 1956.

7 A. M. Freudenthal, "Cumulative Damage Under Random Loading," International Conference on Fatigue of Metals, session 3, paper 4, I. Mech. E. and ASME, 1956.

8 E. Gassner, "Effect of Variable Load and Cumulative Damage on Fatigue in Vehicle and Airplane Structures," International Conference on Fatigue of Metals, session 3, paper 10, I. Mech. E. and ASME, 1956.

9 W. M. Wilson and A. B. Wilder, "Fatigue Tests of Butt Welds in Structural Steel Plates," Bull. 310, University of Illinois Engineering Experiment Station, 1938.

10 H. J. Gough, "The Fatigue of Metals," London, England 1924.

11 B. F. Langer, "Design Values for Thermal Stresses in Ductile Materials," *Welding Journal Research Supplement*, September 1958, pp. 411–417.

Effect of Relaxation on the Behavior of Materials Under Combined Alternating and Static Stress

F. H. VITOVEC

Associate Professor of Metallurgical Engineering, Department of Mining and Metallurgy, University of Wisconsin, Madison, Wis.

The effect of temperature on the stress range diagram is discussed and the particular characteristics of the curves for unnotched and notched specimens are analyzed. Excluding metallurgical factors from consideration, it is suggested that relaxation is the principal mechanism which influences the behavior of polycrystalline metals under combined alternating and static stress. Two parallel nonlinear Maxwell units are used to represent the relaxation mechanism at elevated temperature. An analysis of this model shows that relaxation occurs to an asymptotic finite value of stress which is a function of the initial stress. The same model is applied for representing the behavior in the stress range diagram with the assumption that a linear relationship exists at low temperature between the fatigue strength and the normal stress acting on the planes of reversed slip.

Introduction

THE RAPID DEVELOPMENT of flight vehicles in recent years has led to a demand for more intensified study of the fatigue behavior of materials at both high and low temperatures. Although most of the fatigue data obtained are commonly comparisons of the performance of rather complex materials under some closely specified conditions, comparison of the data permits observation of some apparently basic patterns with regard to effect of temperature and stress concentration as expressed by the stress range diagram.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 21, 1958. Paper No. 59—Met-3.

The effect of temperature on the behavior of heat resistant alloys under combined alternating and static stress is schematically illustrated by the stress range diagrams for unnotched and notched specimens of Fig. 1. The curves represent the behavior of materials in temperature ranges where metallurgical reactions are assumed to be negligible and for stresses which result in a relatively long life for the specimens. Two types of stress range diagrams are shown for both unnotched and notched specimens. For the diagrams to the left in Fig. 1 the ordinate indicates the alternating stress, and the abscissa the mean stress. Each curve gives the combination of alternating and mean stress which results in failure after a definite number of hours or cycles. Dimensionless stress range diagrams are shown at the right in Fig. 1. In these diagrams the ratio of mean stress to static creep rupture stress for the same life is plotted versus the ratio of alternating stress to reversed stress fatigue strength for the same life. The curves shown in both types of diagrams correspond with

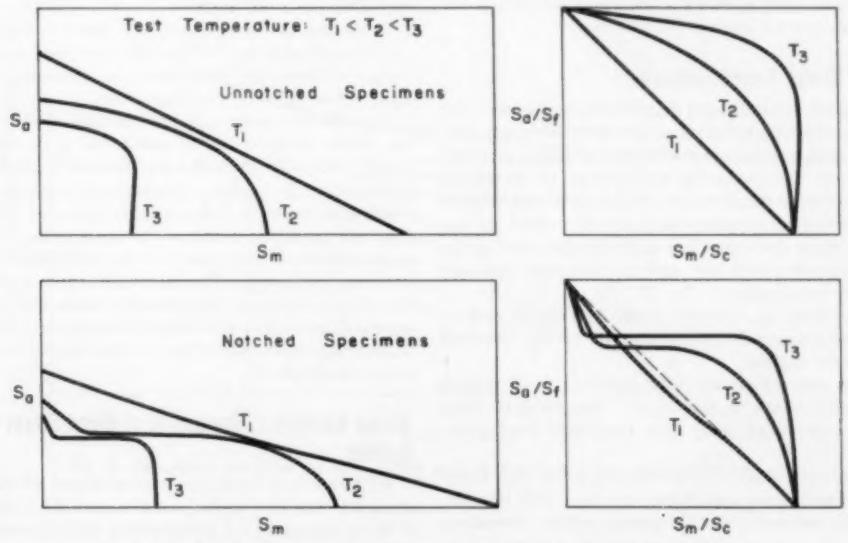


Fig. 1 Schematic stress range diagrams for unnotched and notched specimens

each other. The three lines in each diagram represent the curves for three different test temperatures, where T_1 is a relatively low temperature, T_2 is one intermediate temperature, and T_3 is assumed to be a high temperature which is close to the temperature range for recrystallization.

The schematic diagram for the unnotched specimens shows a nearly linear relationship between alternating stress and mean stress at low temperature. As the temperature increases the relationship becomes nonlinear. At high temperature the curve shows two ranges, one in which the static mean stress has relatively little influence on the allowable alternating stress amplitude, which is followed by a range in which the strength is primarily determined by the rupture stress and the influence of the superimposed alternating stress is small.

In the dimensionless stress range diagram for unnotched specimens the relationship between alternating and mean stress is also linear for low temperature. With increasing test temperature the curves lie increasingly above the straight line.

The stress range diagram for notched specimens indicates three stages which become more pronounced at elevated temperature. Small superimposed mean stress decreases the allowable alternating stress amplitude significantly in the first stage. In the second stage a further increase of static mean stress from a certain value of mean stress has relatively little effect on the allowable stress amplitude, until in the third stage at large mean stress the stress amplitude decreases more or less rapidly to zero. The rate of decrease of the stress amplitude in the third stage increases with increasing temperature. This trend is more significantly apparent from the dimensionless stress range diagram. At small mean stress the curves lie significantly below the straight line. At intermediate alternating to mean stress ratios a more or less pronounced step can be observed which corresponds with the second stage and the curve changes to positions above the straight line.

The described pattern was exhibited in more or less pronounced form by the alloys S-816, Stellite 31, Waspalloy, Inconel X-550, 16-25-6 Timken, GMR-235, SAE 4340, N-155 [1–5].[1] In many instances, especially for notched specimens, the test data commonly procured were not sufficient to show the various characteristic regions of the stress range diagrams.

In the following paragraphs various mechanisms are discussed which may explain the observed behavior, and a model is suggested which permits analysis of the stress range diagram at elevated temperature under consideration of creep.

Analysis of the Stress Range Diagram

The previous discussion indicated that the effect of mean stress on the alternating stress amplitude decreases with increasing temperature. This effect may be caused by (*a*) increasing recovery from fatigue damage with increasing temperature, (*b*) increasing retardation of slip due to strain rate or frequency effects, and (*c*) relaxation. Metallurgical reactions may strongly affect the behavior; however, these are considered to be superimposed to the mechanisms mentioned previously and are therefore excluded from consideration in this paper.

All three mechanisms, i.e., recovery, strain rate effects, and relaxation, occur simultaneously and contribute to the observed behavior in different degrees.

Several studies were concerned with recovery from fatigue damage during intermittent heating [6, 7]. According to these investigations recovery may occur only very early during the fatigue life.

Strain rate effects introduced by alternating stress may cause retardation of dynamic creep and stress rupture. This retardation increases with increasing test frequency and/or increasing temperature [8]. However, experimental studies and mathematical analysis indicate that this mechanism does not cause a large enough effect to explain the behavior as discussed in the introduction.

Relaxation in unnotched specimens may be explained by the following mechanism. In the large cycle fatigue range of polycrystalline specimens fatigue occurs in few localized regions, i.e., individual grains which are subjected to reversed shear. Reversed shear produces "fine slip" and initiation of cracks which spread, join, and produce the fatigue fracture. These fatigue affected grains are embedded in the elastic matrix of the grains which do not participate in reversed slip. The stress range diagrams of polycrystalline metals and tests with single crystals indicate that normal stress, i.e., tensile stress acting on the planes of reversed slip, reduces the fatigue strength. Stress range diagrams for low temperature suggest a linear relationship between reversed shear stress and normal stress.[2] If a mean stress is applied at elevated temperature, relaxation from this normal stress may occur for the grains which are subjected to reversed slip, whereas the elastic matrix carries the static load without significant increase in deformation or change in creep behavior. The localized relaxation of normal stress, however, increases the life of the specimen or, in other words, permits higher stress amplitudes for the same nominal mean stress and the same specimen life.

Localized relaxation explains also the three stages of the stress range diagram for notched specimens [9]. The apparent initial rapid decrease of allowable alternating stress in the first range is due to the fact that nominal stresses (load divided by minimum cross-sectional area of the notched specimen) are plotted instead of true stress (approximately, stress concentration factor times nominal stress). This procedure distorts the diagram and reduces the scale as compared with the curve for unnotched specimens. A further increase in nominal mean stress beyond the corresponding yield strength causes primarily a propagation of the plastic zone and little or no increase in actual mean stress (eventually caused by strain hardening). In other words, relaxation of normal stress occurs in the fatigue affected zone due to plastic deformation and the actual mean stress is nearly constant although the nominal mean stress is further increased. This explains why in the second stage of the stress range diagram the alternating stress amplitude is apparently little or not at all affected by the mean stress.

In the transition to the third stage a further decrease of alternating stress is caused by a change in fracture mechanism from fatigue to creep, and crest stress, i.e., the sum of alternating and mean stress, is the dominating factor.

The effective stress concentration factor which is derived from the stress range data for unnotched and notched specimens usually increases at first with decreasing alternating to mean stress ratio. It reaches a maximum at intermediate ratios from which it decreases to values smaller than one. This trend results from the particular shape of the stress range curve for notched specimens which is produced by the mechanisms discussed in the previous paragraph. The maximum effective stress concentration factor occurs at the transition from the first stage to the second stage of the curve which is determined by the initiation of macroscopic relaxation effects in the volume which is subjected to maximum strain.

Model Analysis of Relaxation of Mean Stress During Fatigue Cycling

With regard to the response to combined alternating and mean stress one may distinguish two phases in the test specimen. One of them represents the grains which are subjected to fatigue and

[1] Numbers in brackets designate References at end of paper.

[2] For a discussion of this relationship see Refs. [11] and [12].

which extend only over a small fraction of the entire volume, the other represents the matrix which is, at elevated temperature, primarily subjected to creep under constant stress. These two phases may be represented by two nonlinear Maxwell units in parallel, which are coupled so that their total strain is the same. This condition is expressed by the equation

$$e_1 + e_{c1} = e_2 + e_{c2} \tag{1}$$

where e is the elastic strain and e_c is the creep strain. Index 1 indicates the unit which represents the fatigue affected grains and index 2 indicates the unit which represents the matrix. Differentiating Equation (1) and assuming that the contribution of phase 1 to the load carrying capacity of the specimen is negligible results in the equation for the strain rates

$$\dot{e}_{c2} - \dot{e}_{c1} = \dot{e}_1$$

or, in another form, if the strain rate is expressed as a function of stress $\dot{e}_c = f(s)$ the differential equation for this model is

$$f(s_2) - f(s_1) = \frac{1}{E}\frac{ds_1}{dt} \tag{2}$$

where E is Young's modulus. Expressing the creep function by the hyperbolic sine relationship $\dot{e}_c = c \sinh (s/s_0)$ and integrating Equation (2) gives the time t at which the stress reduces from the initial stress s_i to the stress s

$$t = \frac{1}{E}\int_{s}^{s_i} \frac{ds}{K - c_2 \sinh (s/s_{02})} \tag{3}$$

where $K = f(s_2) = c_2 \sinh (s_i/s_{02}) = \text{const}$ because no relaxation occurs in phase 2. Integration of Equation (3) results in

$$t = \frac{s_{01}}{Ec_1}\frac{1}{\sqrt{1+(K/c)^2}} \ln \frac{\left[\sqrt{1+(K/c)^2} + (e^{s_i/s_{01}} - K/c)\right]\left[\sqrt{1+(K/c)^2} - (e^{s/s_{01}} - K/c)\right]}{\left[\sqrt{1+(K/c)^2} - (e^{s_i/s_{01}} - K/c)\right]\left[\sqrt{1+(K/c)^2} + (e^{s/s_{01}} - K/c)\right]} \tag{4a}$$

$$\text{for} \quad \left|\frac{e^{s/s_0} - (K/c)}{\sqrt{1+(K/c)^2}}\right| < 1$$

and

$$t = \frac{s_{01}}{Ec_1}\frac{1}{\sqrt{1+(K/c)^2}} \ln \frac{\left[\sqrt{1+(K/c)^2} + (e^{s_i/s_{01}} - K/c)\right]\left[(e^{s/s_{01}} - K/c) - \sqrt{1+(K/c)^2}\right]}{\left[(e^{s_i/s_{01}} - K/c) - \sqrt{1+(K/c)^2}\right]\left[\sqrt{1+(K/c)^2} + (e^{s/s_0} - K/c)\right]} \tag{4b}$$

$$\text{for} \quad \left|\frac{e^{s/s_0} - (K/c)}{\sqrt{1+(K/c)^2}}\right| > 1$$

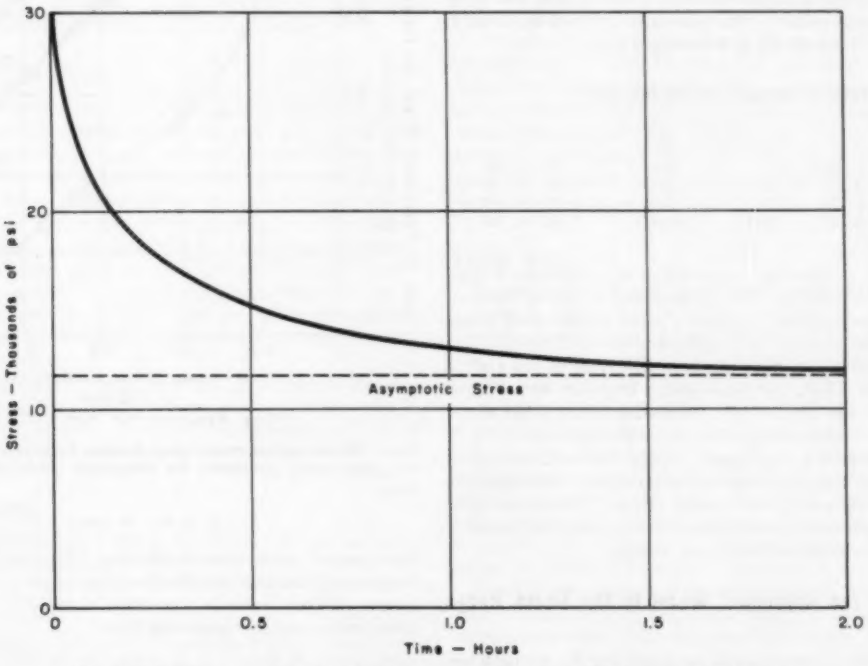


Fig. 2 Relaxation curve for the alloy S-816 at 1350 F and an initial stress of 30,000 psi

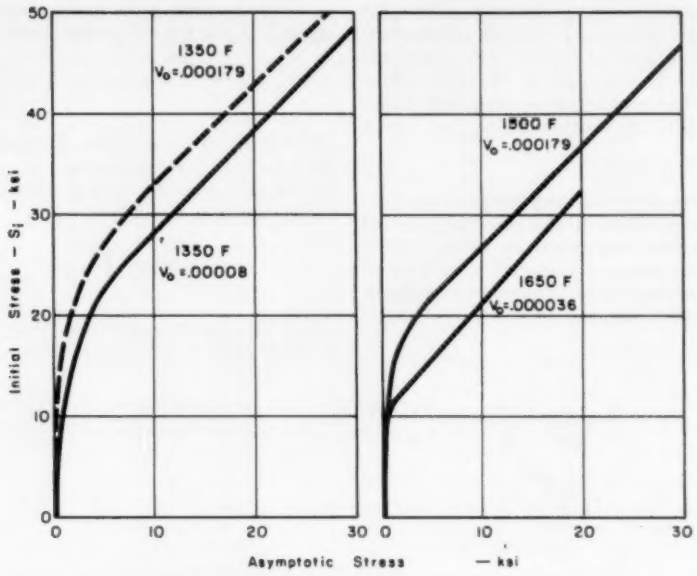


Fig. 3 Relationship between initial stress and asymptotic stress for the alloy S-816 at different temperatures [v_0 corresponds to c_1 in Equation (4)]

Equations (4) were evaluated using data for the alloy S-816 [3]. The constants c_2 and s_{02} were determined from the creep curves using Nadai's chart [10]. Furthermore, it was assumed that the difference in creep behavior between the two phases is characterized by the difference between c_2 and c_1, and that $s_{02} = s_{01}$. The numerial values of the constants is shown in Table 1. The method of determining c_1 is discussed later.

Table 1 Constants for the alloy S-816

Test temperature, deg F	c_2, in/in/hr	s_{02}, psi	c_1, in/in/hr
1350	2.90×10^{-6}	5550	8.00×10^{-5}
1500	1.08×10^{-6}	3300	1.79×10^{-4}
1650	4.00×10^{-6}	1560	3.60×10^{-5}

Fig. 2 shows a relaxation curve which was determined from Equation (4) for S-816 at 1350 F, assuming an initial stress of 30,000 psi. The diagram indicates a rapid relaxation of stress to an asymptotic value s_p = 11,675 psi in about 2 hours. This asymptotic stress is determined by a critical value of $\exp(s/s_0)$ in Equation (4) at which the denominator becomes zero. It increases more or less linearly with increasing initial stress above a certain value of initial stress as shown by the diagrams of Fig. 3. This value of initial stress appears to decrease with increasing temperature, and the rate of increase of asymptotic stress appears to be about the same at all three temperatures. The initial rapid decrease of stress and the occurrence of an asymptotic stress is significant for consideration of fatigue damage.

Application of the Relaxation Model to the Stress Range Diagram

The application of this concept for describing the strength behavior under nonlinear alternating and static stress may be illustrated by use of functional relationships. In a first approximation it is assumed that the relationship between alternating stress s_a and mean stress s_m also is linear at elevated temperature if no relaxation would occur, i.e.,

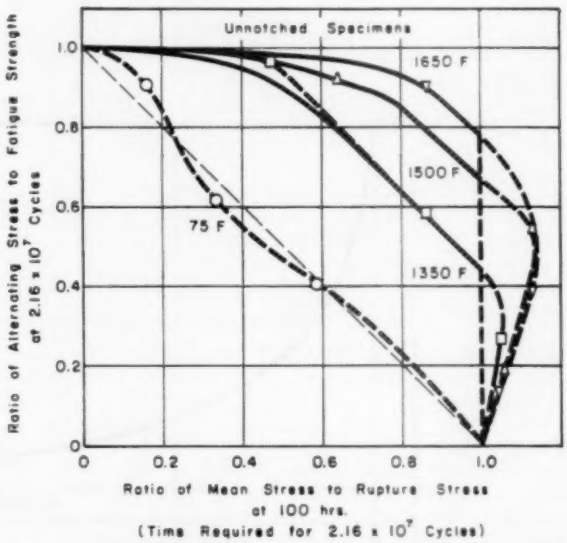


Fig. 4 Dimensionless stress range diagram for unnotched specimens of the alloy S-816 comparing the theoretical curve with experimental data

$$s_a + k s_m = \text{const} \tag{5a}$$

The nominal mean stress is identical with the initial stress s_i. Because of relaxation the effective mean stress s_{meff} acting on the fatigue affected grains is a function of nominal mean stress or initial stress, and decreases with time t

$$s_{meff} = f(s_m, t). \tag{5b}$$

With the relaxation model given this relationship is determined by Equations (4). Since the initial rate of relaxation is very high it may be permitted to assume that fatigue damage corresponds to that produced by a stress equivalent to the relaxed or asymp-

totic stress. Combination of Equations (5a) and (5b) result in

$$\sigma_{eff} = \text{const} - k \cdot f(\sigma_m, t). \quad (5c)$$

Test data for S-816 alloy were evaluated using Equations (4) according to the relationship (5c). The constant c_1 as listed in Table 1 is determined by fitting the theoretical curve through the experimental points.

The theoretical curves are shown in the stress range diagram of Fig. 4 together with the test points. Up to a ratio of mean stress versus creep rupture stress of 1.0 both curves can be brought to close agreement. The theoretical curve is limited by a vertical line at the ratio 1.0 which represents creep rupture. Since the model was derived for the effect of mean stress on fatigue strength only it can not explain the details in behavior in the creep rupture region. The experimental curves indicate in this region that adding a small alternating stress allows a higher mean stress for the same life than static load alone. Considerable evidence points toward retardation of creep rupture caused by strain rate effects which are introduced by alternating stress. A similar retardation effect is observed for dynamic creep and has been analyzed using this concept in a previous publication [8]. A detailed analysis of this effect is beyond the scope of this paper.

Stress Relaxation During Fatigue in Notched Specimens at Elevated Temperature

The concept discussed in the previous paragraph can also be applied to notched specimens for stresses within the first stage of the stress range diagram. In this range fatigue occurs only in few grains along the circumference of the root of the notch. At the initiation of the second stage, however, a relatively large volume is subjected to relaxation effects. There are not enough data available at this time for the first and second stage of the stress range diagram of notched specimens which would permit a theoretical analysis of the mechanism in these ranges.

Summary and Conclusions

If a material is subjected to combined alternating and static stresses relaxation of static mean stress for the grains which are subjected to fatigue may occur primarily at elevated temperatures. This mechanism can be represented by two parallel nonlinear Maxwell units. The model can describe the observed behavior quantitatively and permits the computation of the fatigue stress range diagram. The model analysis indicates that relaxation from mean stress occurs asymptotic to a finite stress which is determined by the initial stress. Other mechanisms such as recovery or strain rate effects which may also affect the behavior are considered to be of secondary importance as compared with the effect of relaxation.

Relaxation effects produce three stages in the stress range diagrams of notched specimens. At present these stages can be analyzed only qualitatively and more experimental information is necessary for a quantitative theoretical analysis.

Acknowledgments

The author wishes to thank J. C. Storry for the numerical evaluation of the data.

References

1 W. J. Trapp, "Elevated Temperature Fatigue Properties of SAE 4340 Steel," Wright Air Development Center, Technical Report 52-325, Part I, Dec., 1952.

2 W. S. Hyler and W. F. Simmons, "Factors Influencing the Notch Fatigue Strengthening of N-115 Alloy at Elevated Temperatures," TRANS. ASME, vol. 78, 1956, pp. 339–347.

3 F. Vitovec and B. J. Lazan, "Creep, Rupture, and Notch Sensitivity Properties of S-816 Alloy Up to 1650 F Under Fatigue and Static Stress," Symposium on Metallic Materials for Service at Temperatures Above 1600 F, Am. Soc. Testing Materials 1955, pp. 69–88. (Issued as separate publication ASTM STP No. 174.)

4 F. Vitovec and B. J. Lazan, "Axial Stress Fatigue, Creep, and Rupture Properties of Unnotched and Notched Specimens of Heat-Resistant Alloys," Papers on Metals, Am. Soc. Testing Materials, 1957, pp. 45–60. (Issued as separate publication ASTM STP No. 196.)

5 F. Vitovec, "Fatigue, Creep, and Rupture Properties of the Alloys Udimet 500, Hastelloy R-235, and GMR-235," Wright Air Development Center, Technical Report 58-340, July, 1958.

6 K. Laute and G. Sachs, "Was ist Ermüdung?" *Zeitschrift, VDI*, vol. 72, 1928, pp. 1188–1189.

7 G. M. Sinclair and T. J. Dolan, "Use of a Recrystallization Method to Study the Nature of Damage in Fatigue of Metals," Technical Report No. 21 of the Behavior of Materials Under Repeated Stress, Contract N6ori-71, T. O. 4, February, 1951.

8 F. Vitovec, "Dynamic Creep With Special Consideration of Strain Rate Effects," *Proceedings, ASTM*, vol. 57, 1957, pp. 977–986.

9 F. Vitovec, "Characteristic Behavior of Heat Resistant Alloys Under Endurance Loads," Berg- u. Hüttenmännische Monatshefte, vol. 102, 1957, pp. 297–303.

10 A. Nadai and P. G. McVetty, "Hyperbolic Sine Chart for Estimating Working Stresses of Alloys at Elevated Temperatures," *Proceedings, ASTM*, vol. 43, 1934, p. 735.

11 W. T. Chodorowski, "Fatigue Strength in Shear of an Alloy Steel With Particular Reference to the Effect of Mean Stress and Directional Properties," *Proceedings, I. Mech. E.*, International Conference on Fatigue of Metals, London, England. September, 1956, pp. 122–131.

12 F. B. Stulen and H. N. Cummings, "A Failure Criterion for Multi-Axial Fatigue Stresses," *Proceedings, ASTM*, vol. 54, 1954, pp. 822–835.

DISCUSSION

M. J. Manjoine[3]

The author points out that at low mean-to-alternating stress ratios the strain rate effects are small compared to the relaxation effects for materials under combined alternating and static stress. At high mean-to-alternating stress ratios, however, net flow occurs and strain-rate effects must be considered. The writer[4] has postulated the rate of loading from the minimum to maximum stress (alternating-stress range) and the maximum stress governs the incremental flow which adds to the creep strain resulting from the *minimum* stress. The effective stress for flow and rupture, therefore, lies between the minimum and mean stress; this results in a higher apparent creep-rupture strength at high mean-to-alternating stress ratios.

Volker Weiss[5]

We are currently conducting thermal-cycling fatigue tests under a research contract with the Army Ballistic Missile Agency. The temperature range of thermal cycling is between −320 F and +600 F. Various light metal alloys such as aluminum and magnesium base alloys are being studied. The specimen is subjected to a constant load and immersed in liquid nitrogen. Thermal cycling is achieved by periodically passing a high current (up to 300 A) through the specimen. Thus maximum temperature, the heating rate, and the duration of the cooling cycle could be varied. The total number of cycles which a specimen withstood for a given set of conditions was recorded as a function of applied nominal stress. Time versus elongation and temperature versus elongation curves are recorded for each test.

Fig. 5 is a photograph of the test stand. The grips of the stress-rupture machine were modified to allow immersing the specimen in liquid nitrogen. The three instruments are (*a*) a Brown con-

[3] Research Engineer, Westinghouse Electric Corporation Research Laboratories, Pittsburgh, Pa. Mem. ASME.

[4] M. J. Manjoine, "Effect of Pulsating Loads on the Creep Characteristics of Aluminum Alloy 14S-T," *Proceedings, ASTM*, vol. 49, 1949, p. 788.

[5] Syracuse University, Syracuse, N. Y.

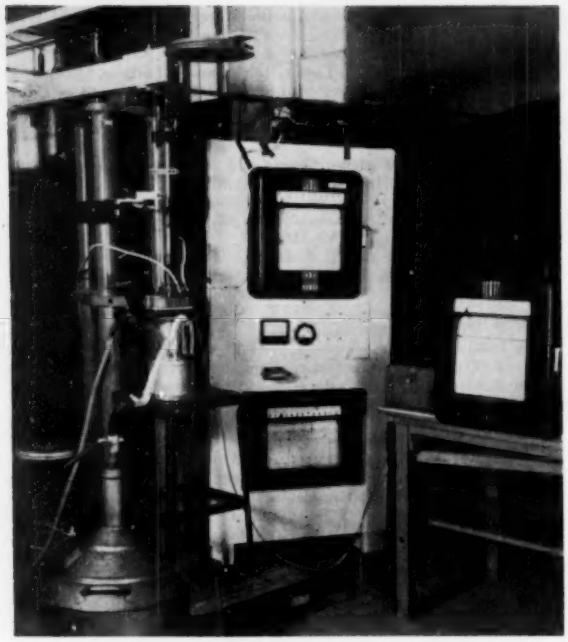


Fig. 5

troller which controls the maximum temperature of each cycle, (*b*) a speedomax recorder which plots the variation in specimen length against time, and (*c*) a speedomax *x-y* recorder which plots instantaneous temperature against length of the specimen.

Fig. 6 shows the effect of applied nominal stress, maximum cycling temperature, and heating rate on the number of thermal cycles withstood to failure of the ALCOA aluminum alloy X-2020. The yield strength values for the respective maximum cycle temperatures after $^1/_2$ hour at temperature, as reported by ALCOA, are also listed for comparison. Specimens were tested under moderate and high heating rate, i.e., approximately 100 F/sec and 1000 F/sec, respectively. It can be seen from Fig. 6 that the maximum cycle temperature appears to be the dominant variable controlling the mechanical behavior of the material under investigation.

Analysis of our data is by no means completed. From the few experimental results shown here, however, it is evident that both maximum cycling temperature and, to a lesser degree, heating rate have an effect on the thermal-cycling fatigue behavior of the material investigated. Both thermal stresses as well as creep seem to be active in promoting failure.

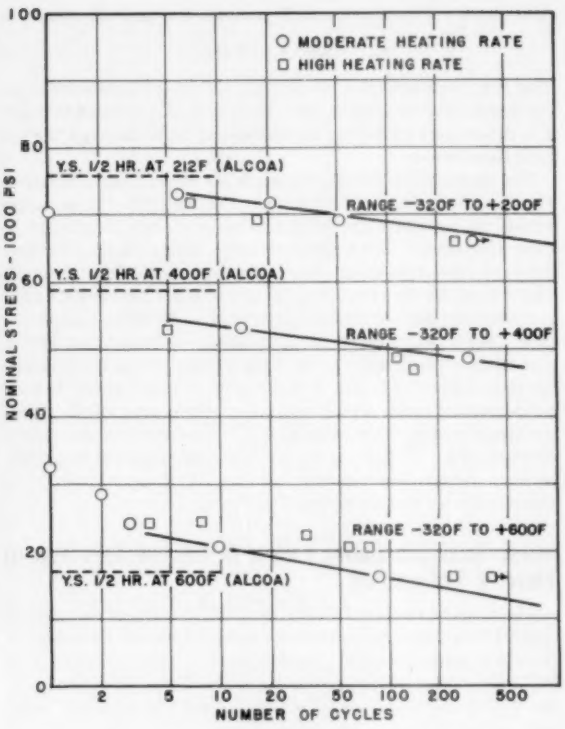


Fig. 6

Author's Closure

Mr. Manjoine draws attention to the creep retarding effect caused by strain rate which is introduced by the superimposed alternating stress. From this behavior it may be inferred that rupture may also be retarded. The effect of strain rate becomes particularly apparent at high alternating-to-mean stress ratios. It may aid in the explanation for the transition from fatigue to stress rupture conditions as indicated in Fig. 4. However, as noted in the paper (page 442) the significant retardation of rupture which is observed in many instances [8] cannot be entirely explained by strain-rate effects.

The interesting study of the effect of mean stress on thermal fatigue under different heating rates reported by Mr. Weiss indicates that probably localized relaxation may influence the strength behavior under the particular test conditions.

Cyclic Operation of Pressure Piping With γ Heating

K. R. MERCKX
Hanford Atomic Products Operation, Richland, Wash.

An elastic-plastic analysis is developed for an internally cooled pressure tube with uniform heat generation. This analysis extends the method of calculating the location of the elastic-plastic boundary reported by Barrie [4][1] to account for the change in the plastic zone due to residual stresses which occur during cyclic operation. Numerical calculations are made for operating conditions expected to be encountered in a pressure tube in a loop through the Engineering Test Reactor Core. The numerical results show that the radius of the initial plastic boundary decreases during subsequent loading cycles. Also, for equal maximum pressure and volumetric heat generation, the total plastic strain per operational cycle on the inner tube surface and the residual tensile stress on the outer tube surface increase when the tube wall is thickened.

Introduction

IN-REACTOR PRESSURE PIPING can undergo gamma heating which causes mechanical effects comparable to those created by the pressure loading. Such a pressure tube is installed in a core facility of the Engineering Test Reactor at Arco, Idaho. The gamma heating in this facility is severe enough to cause plastic yielding for any tube wall thickness sufficient to withstand the internal pressure. Under such operating conditions, design analyses should consider the plastic yielding due to cyclic loading as well as the long term effects of steady-state operation. Hence, the following approximate elastic-plastic solution is developed in order to evaluate the changes in cyclic plastic strains which occur for various pipe thicknesses under the cyclic loads of an internal pressure and a uniform volumetric heat generation.

Methods of Analysis

The analysis of an internally cooled and pressurized pipe with uniform heat generation described in this report is made in order to estimate the elastic-plastic boundaries which occur after residual stresses are established by cycling both the internal pressure and amount of heat generation. The "heating cycle" is considered to be full internal pressure and gamma heat generation and the "cooling cycle" is considered to be zero internal pressure and heat generation. The maximum internal pressure in the pipe is P; the pipe is internally cooled; and the maximum rate of volumetric heat generation is q. The following assumptions have been made regarding the behavior of the material:

1 Perfect plastic-elastic behavior
2 Equal yield strengths in tension and compression
3 No Bauschinger effects [1]
4 The material properties are independent of temperature
5 Maximum stress is the yield criterion (σ_y = yield stress)

In this problem, the maximum stress amplitude in the region of yielding is the circumferential stress, therefore, $\sigma_i = \pm\sigma_y$ in the region of plastic yielding.

In order to facilitate the calculations used in this analysis, the following elastic solutions are used to determine the stress distributions in the elastic regions of the pressure pipe or thick-walled cylinder.

1 The ratio of the circumferential stress [2] to the yield stress $s_T = \sigma_i/\sigma_y$, for an infinite length internally cooled cylinder with uniform volumetric heat generation q and zero surface traction at radii au_{pi}, and au_0 is

$$s_T(u, u_{pi}) = \frac{B}{2u^2}\left\{\frac{2u_0^2}{u_0^2 - u_{pi}^2}\left[(u^2 + u_{pi}^2)u_0^2 \ln\frac{u_0^2}{u_{pi}^2} - (u_0^2 - u_{pi}^2)u^2 \ln\frac{u^2}{u_{pi}^2}\right] + 3u^4 - 5u_0^2u^2 - u_{pi}^2u^2 - u_0^2u_{pi}^2\right\} \quad (1)$$

[1] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 30, 1959. Paper No. 59—Met-4.

Nomenclature

q = internal heat generation/volume
P = pressure
E = elastic modulus
α = coefficient of linear thermal expansion
K = thermal conductivity
v = Poisson's ratio
σ_y = yield stress
$B = \frac{E\alpha q a^2}{8K(1-v)\sigma_y}$, problem parameter containing material properties, heat generation, and size
a = inner tube radius
t = tube thickness
ΔT = temperature drop across the tube
$d = P/\sigma_y$, pressure parameter
u = (radial distance)/(inner radius)
u_0 = (outer radius)/(inner radius)
u_{pi} = (radial distance to plastic boundary)/(inner radius); for cycle i
σ_t = circumferential stress
σ_r = radial stress
$(\ )_i$ = subscript signifying cycle; $(\ )_0$ initial heating; $(\ )_1$ first cooling; etc.
s_i = (total circumferential stress)/(yield stress); for cycle i
$s_r(u_{pi})$ = σ_r/σ_y for the stress in a cylinder with inner pressure P and outer radial stress σ_r if the circumferential stress = σ_y
$s_T(u, u_{pi})$ = σ_t/σ_y for location u in a cylinder with inner boundary u_{pi} for thermal stresses
$s_P(u, u_{pi})$ = σ_t/σ_y for location u in a cylinder with inner boundary u_{pi} for radial stress $s_r(u_{pi})$

where

$$B = \frac{E\alpha q a^2}{8K(1-\nu)\sigma_y}.$$

[1] The ratio of the circumferential stress [3] to the yield stress $s_P = \sigma_t/\sigma_y$ for an infinite length thick-walled cylinder with a surface traction of $s_r = \sigma_r/\sigma_y$ at radius au_{pi} and zero surface traction at radius au_0 is

$$s_P(u, u_{pi}) = -s_r\left(\frac{u_{pi}^2}{u_0^2 - u_{pi}^2}\right)\left(1 + \frac{u_0^2}{u^2}\right) \qquad (2)$$

Both of these equations are for conditions of plain strain where the resultant axial force can be adjusted to any prescribed value with the addition of a uniform axial stress. The equations are written in dimensionless form suitable for numerical evaluation.

The highest stress magnitude in the pipe during the pressurization and heating are the circumferential stresses on and near the inner boundary of the tube; thus, the maximum stresses, the circumferential stresses, reach the yield stress σ_y in the plastic region $1 \leq u \leq u_{pi}$ $(i = 0, 2, \ldots.)$ during the "heating cycle." Since the maximum yielding occurs along the inner surface of the tube, the circumferential residual stresses will be of the highest absolute magnitude along the inner surface of the tube. If the pressure and thermal incompatibilities are of sufficient magnitude, their removal will cause yielding along the inner regions of the tube $1 \leq u \leq u_{pi}$ $(i = 1, 3, \ldots.)$ and the circumferential stress will be $\sigma_t = -\sigma_y$. Therefore, in the plastic portion of the tube, the ratio of the circumferential to yield strength s_i is

$$s_i(u) = (-1)^i, \quad 1 \leq u \leq u_{pi} \qquad (3)$$

The calculations which follow are made by assuming that the change in circumferential stresses within the plastic region can be determined with a knowledge of the previous stress state and the new circumferential stress value. If the change in the circumferential stresses are known then, by equilibrium considerations within the plastic region, the change in radial stress can be determined. The change in radial stress at the elastic-plastic boundary is calculated and assumed to be the surface traction on the elastic portion. The changes in the circumferential stresses in the elastic region then occur due to the change of the temperature within the elastic region and the change of inner surface traction. These changes are either subtracted or added to the original stress system so that the yield conditions are satisfied in the plastic region, thermal effects are accounted for, and the correct pressure exists on the inner tube surface. The location of the elastic-plastic boundary is found by equating the change in the circumferential stress in the elastic region to that of the plastic region at the elastic-plastic boundary. The detailed explanation of this method of determining the location of the plastic boundaries u_{pi} and the total circumferential stresses throughout the tube follows:

Initial Heating and Pressurization

For a pipe in a stress free state, the initial application of pressure and γ-heating (if of sufficient magnitude) will cause plastic yielding on the inner boundary. The circumferential stress distribution is of the form shown in Fig. 1. The force equilibrium equation in y-direction for the plastic portion of the quadrant of pipe shown in Fig. 1 is

$$aP + c\sigma_r = (c - a)\sigma_y \text{ or } s_r(u_{po}) = -\frac{d}{u_{po}} + \left(1 - \frac{1}{u_{po}}\right) \qquad (4)$$

where

$$u_{po} = \frac{c}{a}; \quad d = \frac{P}{\sigma_y}; \quad s_r(u_{po}) = \frac{\sigma_r}{\sigma_y}.$$

The maximum stress criterion specifies the circumferential stress in the plastic region, therefore, allowing equation (4) to be evaluated using only equilibrium considerations. The value of $s_r(u_{po})$, equation (4), is needed to determine the portion of the circumferential stresses in the elastic region due to the surface traction transmitted across the elastic-plastic boundary. The problem in the elastic region now resolves into an internally cooled cylinder with an inner radius au_{po}, uniform internal heat generation q, and the surface traction. Therefore, the total circumferential stress in the elastic region is obtained by superposition of the elastic solutions, equations (1) and (2). At the elastic-plastic boundary, the total circumferential stress equals the yield stress; hence,

$$1 = s_T(u_{po}, u_{po}) + s_P(u_{po}, u_{po}) \qquad (5)$$

where equations (1), (2), and (4) are used to evaluate s_T, s_P, s_r with $u = u_{po}$ and $u_{pi} = u_{po}$. Equation (5) is satisfied only if the correct value of u_{po} is assumed; therefore, equation (5) is solved by trial and error for u_{po}, the first or initial plastic boundary. The resulting stress distribution throughout the cylinder or tube can then be evaluated:

$$\begin{aligned} s_0(u) &= 1 \ldots\ldots 1 \leq u \leq u_{po} \\ &= s_P(u, u_{po}) + s_T(u, u_{po}) \ldots\ldots u_{po} \leq u \leq u_0 \end{aligned} \qquad (6)$$

This method of evaluation was previously used by Barrie [4]. The method of estimating the subsequent stress distributions during cyclic operation is the author's extension of Barrie's method.

Cooling Cycle

When the pressure and heat generation are removed, the residual circumferential stress is found by subtracting the elastic solutions

$$s_P(u, u_{pi}) \text{ and } s_T(u, u_{pi}) \text{ from } s_{i-1} \; (i = 1, 3, 5, \ldots)$$

in all of the cylinder which behaves elastically. If conditions are severe enough, plastic yielding in compression occurs on the inner portion of the cylinder. The radius of the elastic-plastic boundary during cooling will be less than that of the preceeding heating cycle; hence, throughout the new plastic region and on its boundary, the decrease in the circumferential stress is twice the yield stress. The new value of s_r used in the calculations for the stress decrease within the elastic region is determined with the same equilibrium considerations as were previously used in obtaining equation (4), except σ_y is replaced by $2\sigma_y$. The result of this analysis is

$$s_r(u_{p1}) = -\frac{d}{u_{p1}} + 2\left(1 - \frac{1}{u_{p1}}\right). \qquad (7)$$

The new elastic-plastic boundary is determined in a manner similar to that described for equation (5), except the decrease of

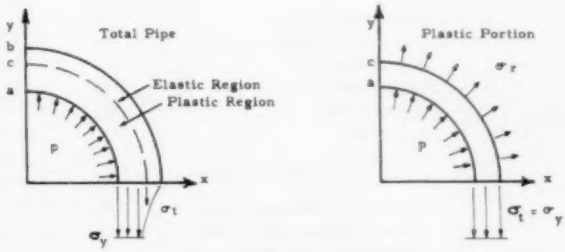


Fig. 1 Circumferential stress distribution

the total circumferential stress in the elastic region at the new elastic-plastic boundary equals twice the yield stress:

$$2 = s_T(u_{p1},u_{p1}) + s_P(u_{p1},u_{p1}). \quad (8)$$

Equation (8) is solved by trial and error for the new elastic-plastic boundary u_{p1}. The u_{pi} ($i = 1, 3, 5, \ldots$) remains constant since all radii for the cooling plastic boundaries are less than those for the heating plastic boundaries in this analysis; hence, equations (7) and (8) do not change for the different cooling cycles. The ratio of the total circumferential stress to yield stress for any cooling cycle is:

$$\begin{aligned} s_i(u) &= -1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1 \le u \le u_{p1} \\ &= s_{i-1}(u) - s_P(u, u_{pi}) - s_T(u, u_{pi}) \ldots u_{p1} \le u \le u_0 \quad (9) \end{aligned}$$

where $i = 1, 3, 5, \ldots$ Equations (1) and (2) are used to evaluate s_P and s_T while s_{i-1} is obtained from the previous heating cycle and $s_r(u_{p1})$ is given by equation (7).

Subsequent Heating Cycles

The plastic yielding occurring during previous cycles of operation causes residual stress systems which affect the amount of increase in circumferential stress necessary for plastic yielding. The radius of the cooling plastic boundary is less than the radius of the subsequent heating plastic boundary; therefore, the increase of the circumferential stress necessary to cause plastic yielding at the elastic-plastic boundary is less than twice the yield stress and is

$$1 - s_{i-1}(u_{pi}) = s_T(u_{pi}, u_{pi}) + s_P(u_{pi}, u_{pi}) \ldots (i = 2, 4, \ldots)$$

where $s_{i-1}(u_{pi}) > -1$.

In order to determine the radial stress used for the surface traction in equation (2) for s_P at the elastic-plastic boundary, only the increase in circumferential stress in the new plastic region is used in the equilibrium argument. Thus the shaded area in Fig. 2 is the force increase due to the change in circumferential stress in the plastic region. Using the same equilibrium considerations as previously used to obtain equation (4),

$$aP + au_{pi}\sigma_r \simeq a(u_{pi} - 1)2\sigma_y - \frac{a(u_{pi} - u_{pi})}{2}(\sigma_y + \sigma_{t_{i-1}})$$

or

$$s_r(u_{pi}) = -\frac{d}{u_{pi}} + 2\left(1 - \frac{1}{u_{pi}}\right) - \frac{(u_{pi} - u_{pi})}{2u_{pi}}(1 + s_{i-1}). \quad (10)$$

The ratio of total circumferential stress to yield stress for any heating cycle other than the initial is

$$\begin{aligned} s_i(u) &= 1 \ldots\ldots\ldots 1 \le u \le u_{pi} \\ &= s_{i-1}(u) + s_T(u, u_{pi}) + s_P(u, u_{pi}) \ldots u_{pi} \le u \le u_b, \quad (11) \end{aligned}$$

where $i = 2, 4, 6 \ldots$. The plastic boundary is determined by substituting a series of increasing u_{pi} into the second part of equation (11). The first trial u_{pi} that satisfies $s_i(u) \le 1$ for all $u \ge u_{pi}$ is the correct u_{pi}. Equations (1), (2), and (10) are used to evaluate s_T, s_P, and s_r.

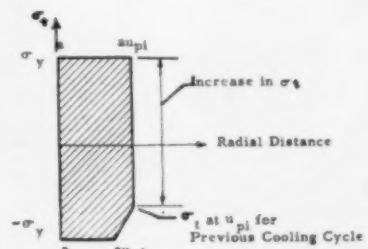


Fig. 2 Change of circumferential stress in the plastic region during heating

Cases for Evaluation

The following cases are evaluated in order to determine the effects of various pipe thicknesses and yield stresses on the total plastic yielding per cycle. The calculations are made for design conditions of a pressure pipe in the Engineering Test Reactor 6 × 9 core facility. The material properties are for 347 stainless steel in the annealed condition and at 600 F. The following properties [5] are used in the numerical calculations:

Design pressure, $P = 2{,}500$ psi
Assumed γ, heating, $q = 30$ w/gm $= 3{,}945$ w/in.^{3}
Inner tube radius, $a = 1.719$ in.
Material, 347 s.s., properties 600 F
Poisson's ratio, $v = 0.28$
Elastic modulus, $E = 21 \times 10^6$ psi
Yield strength, $\sigma_y = 32{,}000$ psi
Thermal conductivity, $K = 11.2$ Btu/hr-F-ft $= 0.273$ w/in-F
Coefficient of thermal expansion, $\alpha = 9.8 \times 10^{-6}$ (in./in.)/F

A thickness was calculated [6] which minimized the total circumferential stress for an elastic material. This calculation yielded a thickness $t = 0.114$ in., a thermal circumferential stress component $(\sigma_t)_T = 18{,}600$ psi, and a circumferential stress due to pressure $(\sigma_t)_P = 39{,}100$ psi. Since the sum of these circumferential stresses is above the yield stress of annealed 347 stainless steel, and the pressure component of stress was above any allowable working stress, the design on an elastic basis is unrealistic. Therefore, the plastic-elastic calculations using the method described in this report are used to estimate actual conditions in thicker pipe sections. Trial thicknesses are selected for reasonable working stresses σ_w calculated on the basis of only considering pressure effect. Thicknesses corresponding to working stresses σ_w are calculated [7] with the equation

$$t = \frac{Pa}{\sigma_w - 0.6P}$$

The following values of working stress are assumed:

$$\sigma_w = 22{,}000;\ 18{,}000;\ 15{,}000 \text{ psi}$$

Using the three values of working stresses and two values of yield stress of 32,000 and 25,000 psi, the following six cases were evaluated by the elastic-plastic solution described in this report:

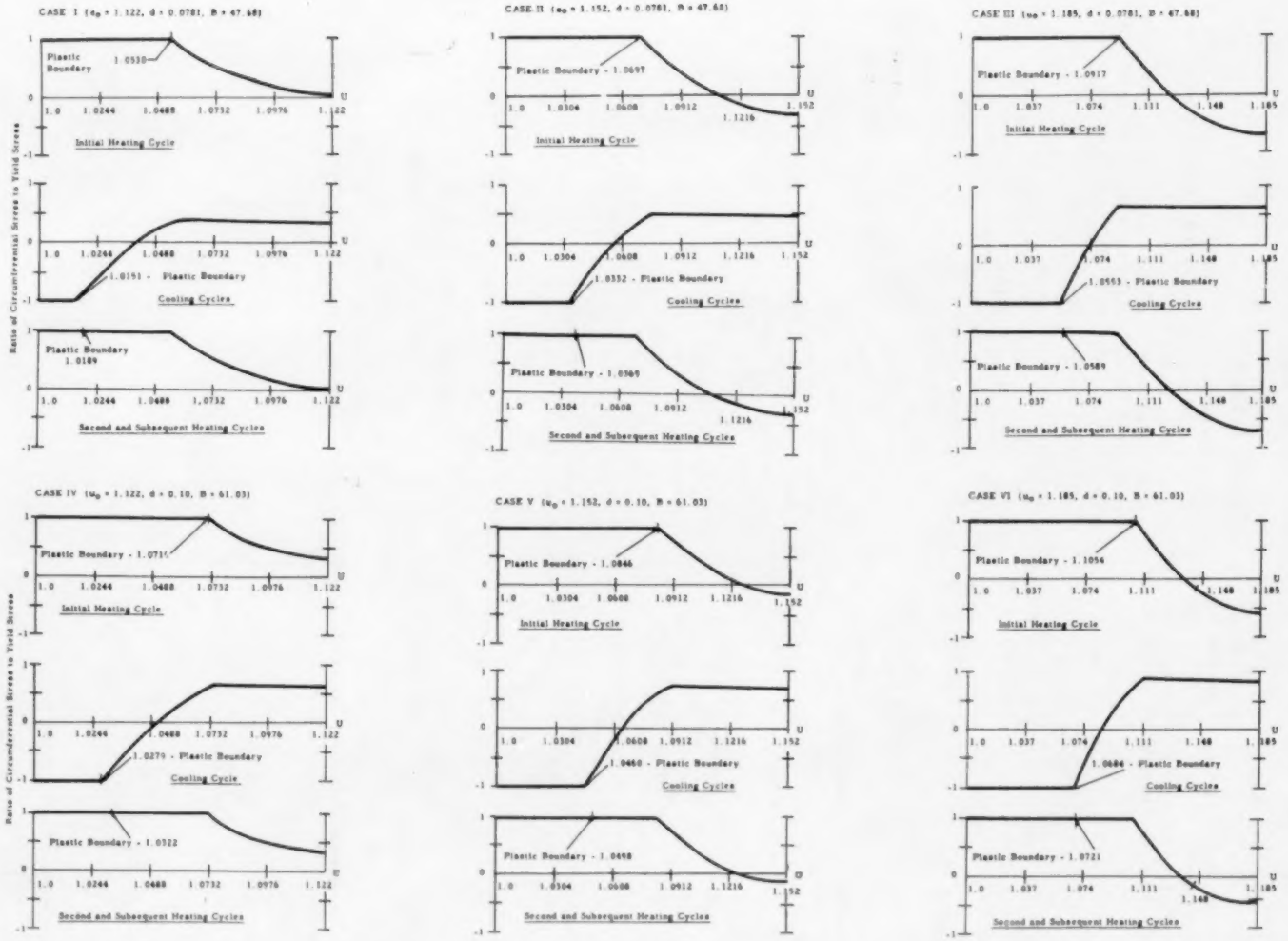
Case	u_0	$d = p/\sigma_y$	$B = \frac{Eaqa^2}{8K(1-v)\sigma_y}$
1	1.122	0.0781	47.68
2	1.152	0.0781	47.68
3	1.185	0.0781	47.68
4	1.122	0.1	61.03
5	1.152	0.1	61.03
6	1.185	0.1	61.03

The temperature drops across the tube ΔT and inner surface heat fluxes are calculated with the following equations:

$$\Delta T = \frac{qa^2}{4K}(u_0^2 \ln u_e^2 - u_0^2 + 1)$$

$$\text{Flux} = \frac{a}{2}(u_0^2 - 1)q$$

The values of these quantities for the various working stresses are given in the following table:


CASE I (u_o = 1.122, d = 0.0781, B = 47.68)
Plastic Boundary
1.0530
Initial Heating Cycle
1.0151 - Plastic Boundary
Cooling Cycles
Plastic Boundary
1.0189
Second and Subsequent Heating Cycles
CASE II (u_o = 1.152, d = 0.0781, B = 47.68)
Plastic Boundary - 1.0697
Initial Heating Cycle
1.0332 - Plastic Boundary
Cooling Cycles
Plastic Boundary - 1.0369
Second and Subsequent Heating Cycles
CASE III (u_o = 1.185, d = 0.0781, B = 47.68)
Plastic Boundary - 1.0917
Initial Heating Cycle
1.0553 - Plastic Boundary
Cooling Cycles
Plastic Boundary - 1.0589
Second and Subsequent Heating Cycles
CASE IV (u_o = 1.122, d = 0.10, B = 61.03)
Plastic Boundary - 1.0715
Initial Heating Cycle
1.0279 - Plastic Boundary
Cooling Cycle
Plastic Boundary - 1.0322
Second and Subsequent Heating Cycles
CASE V (u_o = 1.152, d = 0.10, B = 61.03)
Plastic Boundary - 1.0846
Initial Heating Cycle
1.0460 - Plastic Boundary
Cooling Cycles
Plastic Boundary - 1.0498
Second and Subsequent Heating Cycles
CASE VI (u_o = 1.185, d = 0.10, B = 61.03)
Plastic Boundary - 1.1054
Initial Heating Cycle
1.0604 - Plastic Boundary
Cooling Cycles
Plastic Boundary - 1.0721
Second and Subsequent Heating Cycles
Ratio of Circumferential Stress to Yield Stress
U

u = RADIAL DISTANCE / INNER RADIUS, u_o = OUTER RADIUS/ INNER RADIUS, d = PRESSURE/ YIELD STRESS,

$$B = \frac{E\alpha q a^2}{8K(1-\nu)\sigma_y}$$

Fig. 3 Circumferential stress distributions in γ-heated pressurized pipes

Working stress, psi	Thickness, in.	u_0	Temperature drop, F	Heat flux, Btu/hr-ft^2
22,000	0.210	1.122	331	432,000
18,000	0.261	1.152	517	545,000
15,000	0.318	1.185	774	674,000

Results of Calculations

The numerical results obtained for the total circumferential stresses are graphed in Fig. 3, Case I to Case VI. The plastic boundary becomes larger as the thickness increases or as the yield strength decreases. Subsequent changes in the stress distributions for either the cooling or heating cycles are insignificant, but the radius of the plastic boundary for the first heating cycle is significantly reduced during subsequent heating cycles.

As the pipe thickness is increased or the yield strength is decreased, the residual tensile stress during the cooling cycle on the outer surface of the tube is increased. This stress almost reaches the yield stress for Case VI, which is the example with the thickest pipe wall and the lowest yield stress.

Values for the plastic straining on the inner surface of the tube can be estimated by assuming:

(*a*) No volume change occurs in the plastic region due to the stresses;

(*b*) Axial strain is constant and is known;

(*c*) The axial and radial stresses at the plastic-elastic boundary do not appreciably alter the circumferential strain.

Thus, the displacement on the plastic-elastic boundary, $\delta(u_{pi})$, is:

$$\delta(u_{pi}) \cong a u_{pi} \left[\frac{\sigma_y}{E} + \alpha T(u_{pi}) \right]$$

Accounting for the total change in volume in the plastic region due to the thermal gradients, the following relation for the strain at the inner boundary is obtained:

$$\epsilon_i(1) = \frac{\delta(1)}{a} \cong u_{pi} \frac{\delta(u_{pi})}{a} + (u_{pi}^3 - 1) \frac{\epsilon_z}{2} - \frac{3}{2} \int_1^{u_{pi}^3} \alpha T(u^3) du^3$$

The combination of these two equations gives the following estimate of the circumferential plastic strain:

$$\epsilon_p = \epsilon_T(1) - \alpha T(1) - \frac{\sigma_y}{E} = (u_{pi}^2 - 1) \left(\frac{\sigma_y}{E} + \frac{\epsilon_z}{2} \right) + \alpha u_{pi}^2 T(u_{pi}) - \alpha T(1) - \frac{3}{2} \int_1^{u_{pi}^3} \alpha T(u^3) du^3. \quad (12)$$

For the internally cooled tube with $T(1)$ taken as a reference temperature, α = constant, and $\epsilon_z = 0$, equation (12) becomes

$$\epsilon_p = (u_{pi}^2 - 1) \frac{\sigma_y}{E} \left[1 + (1 - \nu) B \left(3 u_0^2 - \frac{u_0^2 u_{pi}^2 \ln u_{pi}^3}{u_{pi}^3 - 1} - \frac{u_{pi}^3}{2} - \frac{3}{2} \right) \right].$$

The following table contains values for the plastic boundaries and the total plastic strain per cycle calculated with this equation:

Case	Cooling boundary u_{p1}	Heating boundary u_{p2}	Plastic strain per cycle, per cent
I	1.0151	1.0189	0.182
II	1.0332	1.0369	0.449
III	1.0553	1.0589	0.859
IV	1.0279	1.0322	0.301
V	1.0460	1.0498	0.579
VI	1.0684	1.0721	1.014

Acknowledgment

The author would like to thank D. R. Miller for finding an error in my initial method for estimating the plastic strains.

References

1 R. Hill, "The Mathematical Theory of Plasticity," Oxford University Press, London, England, 1950.

2 "Revised Edition Reactor Handbook," Sec. 5.2, p. 18, eq. 31.

3 S. Timoshenko, "Strength of Materials," Part 2, McGraw-Hill Book Company, Inc., New York, N. Y., 1941, p. 239, eq. 202.

4 H. E. Barrie, "An Evaluation of the KAPL 3X3 In-Pile Tube Design," KAPL-M-HEB-1, August, 1957.

5 "ASM Metals Handbook," p. 555.

6 K. R. Merckx, "Pipe Thicknesses Which Minimize Stresses Due to Internal Pressure and Gamma Heating," *Nuclear Science and Engineering*, vol. 3, no. 6, June, 1958, pp. 774–777.

7 ASME Boiler Code, Unfired Pressure Vessels.

DISCUSSION

Irwin Berman[2]

The author is to be commended for his presentation of a practical example in which it is necessary to determine the elastic-plastic behavior of an infinite length thick-walled cylinder subjected to cyclic loading. The problem is solved by the use of the maximum normal-stress yield criterion. As mentioned by the author in his presentation, however, the use of the maximum shear-stress yield condition with its associated flow rule may lead to a better solution.

The maximum shear-stress yield condition with its associated flow rule has previously been employed in similar problems. Prager and Hodge [8][3] and Hodge [9] discussed the thick cylinder in terms of shakedown under repeated internal pressure loading and unloading. Bland [10] presented a more extensive solution in which he considered, "elastoplastic thick-walled tubes of work-hardening materials subject to internal and external pressures and to temperature gradients." The solution for a perfectly plastic material was also discussed by Bland [10] as a special case.

The assumption by the author of a perfectly plastic material may be reasonable even for a work-hardening material such as the 347 stainless steel which was chosen for evaluation. This is so because the approximate plastic-strain-per-cycle values as calculated by the author, for the fairly small ratios of the outer to the inner diameter of the tube, are low. Any extension of this paper to cylinders in which there may be strains of a few per cent would necessitate the consideration of strain-hardening.

This discusser would like to see some graphs of temperature plotted against thickness for the various conditions so as to be better able to evaluate the limitations of the author's assumption that the material properties are independent of temperature.

The discusser believes there may be an error in the values of the elastic modulus and Poisson's ratio which the author used for 347 stainless steel in the annealed condition at 600 F. The values of these material properties are not evident in the author's reference [5].

References

8 W. Prager and P. G. Hodge, Jr., "Theory of Perfectly Plastic Solids," John Wiley & Sons, New York, N. Y., 1951, pp. 110–115.

9 P. G. Hodge, Jr., "Shakedown of Elastic-Plastic Structures," *Residual Stresses in Metals and Metal Construction*, edited by W. R. Osgood, Reinhold Publishing Corporation, New York, N. Y., 1954, pp. 163–191.

10 D. R. Bland, "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and to Temperature Gradients," *Journal of Mechanics and Physics of Solids*, vol. 4, 1956, pp. 209–229.

[2] Consulting Structural Analyst, Research Division, Foster Wheeler Corporation, New York, N. Y. Mem. ASME.

[3] Numbers in brackets designate References at end of this discussion.

D. R. Miller[4]

The problem which the author has analyzed illustrates quantitatively a concept which has been widely understood qualitatively. This concept is that large thermal strains in a ductile metal produce plastic flow rather than extremely high stresses. It is now also widely recognized that plastic thermal strains can produce fatigue failures when the temperatures are cycled a sufficient number of times.

It should be noted that the author has not followed Coffin [11][5] in defining the plastic-strain range as the amount of tensile or compressive plastic strain during a complete cycle of loading, but rather shows in his table the arithmetic sum of the tensile and compressive plastic strains during a cycle, computed by using equation (12).

An interesting observation that may be made in studying the present problem is that the maximum plastic strain which occurs on loading or unloading after the initial application of load is not simply the difference between the maximum strain computed on an elastic basis and the actual elastic-strain range from compressive yielding to tensile yielding. The simple situation, which is characteristic of a temperature-cycled bar with constrained ends, is modified in the tube, because in the latter the occurrence of plastic flow over a portion of the wall thickness produces an expansion or contraction of the tube diameter in a direction which increases the magnitude of plastic strain. These changes in diameter caused a 13 per cent increase in the amount of plastic flow in the writer's approximate analysis of circumferential strains in a thin-walled tube having thermal and membrane stresses similar to those of the author's Case IV. Thus it is evident that elastic-stress analysis alone is not a sufficient basis for evaluation of low-cycle fatigue problems in cases where significant strains can be produced in elastically stressed regions because of plastic flow which occurs elsewhere.

The author has assumed that the thermal and pressure loadings are cycled simultaneously. This assumption is not realistic for an in-pile pressure tube, since the pressure loading will be applied prior to start-up of the reactor and will be removed after shutdown of the reactor. Thus the thermal loading is cycled while the pressure load is maintained. A recent paper by the writer [12] showed that this condition would produce a progressive growth in diameter of the tube if the loadings were sufficiently severe. The criteria derived by the writer indicate that progressive growth would occur in all of the author's cases if the heating were applied and removed repetitively while pressure loading was maintained.

After establishing that adequate dimensional stability is assured in a particular in-pile tube design, a complete evaluation would include an analysis of longitudinal stress and strain distributions. The results of the analysis of circumferential and longitudinal stresses and strains then would be combined to determine the triaxial stress-strain history in the region of maximum plastic strain. If suitable criteria were available, these conditions of low-cycle fatigue under triaxial stresses could be correlated with low-cycle fatigue data on uniaxially stressed specimens. Consideration is being given by various groups to experimental studies directed at correlations between uniaxial and biaxial stress low-cycle fatigue data, but little, if any, correlated data are available. Lacking such correlation, it may be reasonable to assume that a specified range of radial strain in the tube is approximately equivalent to the same longitudinal strain range in a uniaxially stressed low-cycle fatigue specimen of the same material. This assumption attributes primary significance to similarity of triaxial strains in the tube and in the fatigue specimen and ignores the significance of the differences in the stresses.

The author has pointed out the need for consideration of the long term effects of steady-state operation. The most important factor to be considered will probably be the effect of neutron bombardment, which may increase the yield strength of a ductile metal and reduce its ductility. These effects must be considered with respect to low-cycle fatigue and, in some materials, must be considered with respect to the possibility of elevation of the brittle-fracture-transition temperature to an unsafe level.

[4] Consulting Engineer, Knolls Atomic Power Laboratory, Schenectady, N. Y. Mem. ASME.

[5] Numbers in brackets designate References at end of this discussion.

References

11 L. F. Coffin, Jr., "An Investigation of Thermal-Stress Fatigue as Related to High-Temperature Piping Flexibility," TRANS. ASME, vol. 79, 1957, pp. 1637–1651.

12 D. R. Miller, "Thermal-Stress Ratchet Mechanism in Pressure Vessels," TRANS. ASME, series D, JOURNAL OF BASIC ENGINEERING, vol. 81, 1959, pp. 190–196.

Author's Closure

As I. Berman points out, the solution presented in this paper does not use the usual shear stress or Mises yield criterion. Had the author been familiar with Bland's [10] work, the shear stress criteria would have been applied before programming the numerical calculations. After the numerical results were obtained, the author had derived equations which were based on the shear stress yield criterion and these equations would not have presented additional programming difficulties. In the relatively thin-walled tube analyzed in this paper, the variation in numerical results based on the maximum shear stress yield criterion or maximum stress criterion would not be significant. D. R. Miller's comments on critical evaluation of calculated stress and strain conditions in terms of physical behavior when mechanical behavior is not completely understood justify the use of approximate numerical analyses. The purpose of this paper was not to present an elegant method of analytic analysis, but was to illustrate that elastic analyses are not sufficient for calculating the plastic strain per cycle in systems undergoing plastic strain cycling. D. R. Miller's suggestion that thermal-stress ratcheting should also be checked shows that design of structures under severe conditions will require a complete understanding of possible failure mechanisms. The analytical work of the industrial designer is rapidly becoming more complex as service conditions are increased. Both more understanding of material behavior and new calculational methods are required; at present, many practical problems defy analysis because the material behavior cannot be given a realistic representation in terms of the variables which determine their life under actual service conditions.

The assumption of material properties which are independent of temperature is not valid for the thick-walled cylinders with high γ-heating. Temperature drops across the tubes are given in the paper, and, as is listed, the tube with 0.318-in. thick wall would have a 774-F temperature variation. Considering possible film drops and a 550-F coolant temperature an outer tube wall temperature of 1400 F would be reached. At this temperature, elastic analysis ceases to be valid and creep would occur. Thus, the design of thicker tubes does not only cause higher plastic strains based on an elastic-plastic analysis, but the increased temperatures cause the creep phenomenon to be significant. Thus, in the actual selection of a tube design, the analysis present in this paper is just one step in the selection of an adequate tube. A complete analysis would require a creep model which would allow for temperature, irradiation, and load cycling. Since such data are not available, only possible mechanisms of behavior can be estimated by models such as presented in this paper and D. R. Miller's paper [12]. Knowing the inadequacy of the model for material behavior makes numerical values only of relative significance.

E. A. DAVIS
Mechanics Department, Westinghouse Research Laboratories, Pittsburgh, Pa. Mem. ASME

Creep Rupture Tests for Design of High-Pressure Steam Equipment

Creep rupture tests on tubular specimens of type 316 stainless steel were run at 1200 F and at pressures up to 24,000 psi. The specimens were tested under pure internal pressure and equal biaxial tensions. The results of these tests correlate favorably with those of uniaxial tension tests if a comparison is made on the basis of effective stress and effective strain rate.

Introduction

THE DESIGN OF CERTAIN steam-turbine components for advanced high-pressure, high-temperature installations such as the Philadelphia Electric Company's Eddystone No. 1 Unit and the Avon No. 8 Unit of the Cleveland Electric Illuminating Company, demand new investigations into the combined stress behavior of materials to help assure the validity of the design concepts. At supercritical pressures the radial component of stress is of significant magnitude to become a dominant factor in the design of either cylindrical or spherical pressure vessels. With the present day state of design knowledge, it is perplexing to decide what the allowable design stress should be when using triaxial combined stress formulas where all three stress components are of significant values.

To help fill the void of actual test data on combined stress behavior, the author's Company instituted a rather unique testing program in which tubular specimens of type 316 stainless steel were subjected to triaxial stresses of various component ratios. Both pure internal pressure and equal biaxial tension tests were run. The test temperature was 1200 F, and pressures up to 24000 psi were used in making the tests. The ratios of the stress components used in the various tests are shown in Table 1. These ratios represent the average values of the stress components.

This paper suggests a possible means of correlating triaxial test data with data obtained from conventional uniaxial creep tests. It is hoped that a rather thorough analysis of a few complicated tests will shed sufficient light on the triaxial stress creep problem to permit design standards to be established on the basis of uniaxial tension creep tests.

Material

The specimens were cut from a 12-in. length of a $9^3/_4$ OD by $4^5/_8$ ID sleeve forging. This material was reheated after forging to 1950 F and held at this temperature for 2 hours, after which it was quenched in water. The composition and room temperature properties are as follows:

C	Mn	P	S	Si	Ni	Cr	Mo	Cu
0.06	1.28	0.012	0.013	0.38	11.21	17.18	2.25	0.22

Yield strength	31000 psi
Tensile strength	79000 psi
Elongation	78.6 per cent
Reduction in area	80.9 per cent
Vickers hardness	140 DPN

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, March 18, 1959. Paper No. 59—Met-14.

Test Specimens and Testing Conditions

The uniaxial creep tests were run on a 0.505-in. diameter bar with a 3-in. gage length. The tubular specimens had a 6-in. cylindrical gage length. The outside diameter in all cases was $1^1/_8$ inches. The inside diameter of most of the tubes was $^9/_{16}$ in. but for a few it was $^5/_8$ in.

Two types of combined stress loadings, pure internal pressure and equal biaxial tensions, were used. In the equal biaxial tension tests an axial load was added to the internal pressure to make the average axial stress equal to the average tangential stress. Since, under creep conditions, the stresses all vary with the radius the local axial and tangential stresses were not equal.

The tests were run in the testing machine shown in Fig. 1. This photograph shows a specimen in the machine without the furnace. The axial load for the biaxial tension tests was maintained by keeping a constant deflection of the spring weighbar at the lower part of the machine. The internal pressure was supplied by an air-operated pump. The pressure fluctuations were held to a minimum by providing the pressure system with a calibrated leak. By adding a chamber filled with oil at room temperature to the high-pressure system and by providing a leak of one drop of oil about every 5 seconds through a capillary tube a continuous motion of the actuating piston was assured. This reduced the friction effect and kept the driving piston from "freezing" to the cylinder wall. The specimen was connected to this high-pressure system by heavy-walled tubing of small diameter. Thus the high-pressure water at room temperature in the pumping system became high-pressure steam in the specimen inside the furnace.

Table 1

Type of test	σ_θ/σ_z	σ_θ/σ_z	σ_r/σ_z	$\sigma_1:\sigma_2:\sigma_3$
Uniaxial tension, solid cylinders	1	0	0	1:0:0
Pure internal pressure, OD/ID = 2	1	3	$-^1/_2$	$1:^1/_2:-^1/_2$
Biaxial tension, OD/ID = 2	1	1	$-^1/_2$	$1:1:-^1/_2$
Pure internal pressure, OD/ID = 3	1	4	-4	$1:^1/_4:-1$
Biaxial tension, OD/ID = 3	1	1	-1	$1:1:-1$

Special precautions had to be taken for protection against the explosive conditions at fracture. A reinforcing tube of stainless steel was put inside the furnace as a liner to protect the furnace windings. The outside of the furnace was covered with heavy steel tubing.

In the equal biaxial tension tests the axial strain was measured with an extensometer attached to the shoulders of the specimen as shown in Fig. 1. Some of the tests were stopped from time to time and the specimen removed and the increase in diameter measured.

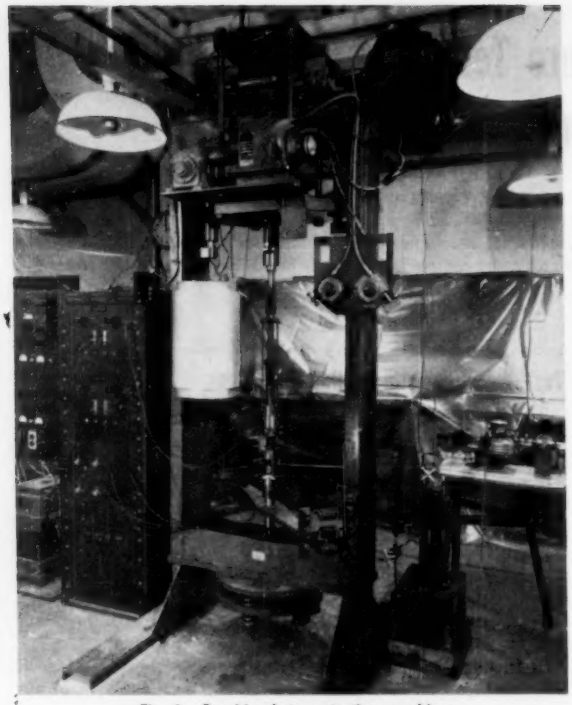


Fig. 1 Combined stress testing machine

These measurements were necessary to determine the distribution of the strains in the tube wall.

In the pure internal pressure tests the increase in diameter was measured by a lateral extensometer with a 10× magnification. The lengths of these specimens were checked before and after test to determine whether or not any creep occurred in the axial direction.

Stress and Strain Distribution

The stress distribution in a heavy-walled cylinder depends upon the strength or creep theory under consideration. In any event progress can only be made by measuring or calculating the strain distribution. The first problem then is to attempt to determine the distribution of the strains or more particularly the distribution of the strain rates. This can be done with a minimum of measurements as long as the material deforms uniformly. Any such analysis will begin to deviate from the actual behavior as soon as an ultimate load is reached in a tension test or when the third stage of creep begins in a creep test.

In a heavy-walled tube under internal pressure the radial and tangential strains at the bore are much larger than those at the outside surface. It appears advisable, then, to evaluate the strain distribution in terms of the natural strains $\bar{\epsilon}$, although it may be necessary later to revert to the conventional strain ϵ.

The strains considered here are the total strains. There has been no attempt to separate the elastic and the plastic components of the strain. After the specimen is loaded in a creep test, however, all the measurable strain is plastic, and so the analysis should be considered to be in terms of plastic strain with the elastic component neglected.

The natural strain components for conditions existing in the present tests ($\epsilon_z \neq f(r)$) can be expressed as

$$\bar{\epsilon}_\theta = \frac{1}{2} \ln \left(\frac{c_1 r^2 + c_2}{r^2} \right) \tag{1}$$

$$\bar{\epsilon}_r = \frac{1}{2} \ln \left(\frac{c_1^2 r^2}{c_1 r^2 + c_2} \right) \tag{2}$$

$$\bar{\epsilon}_z = -\frac{1}{2} \ln c_1^2 \tag{3}$$

where Equation (3) is a definition of the constant c_1 in terms of the known or assumed strain in the axial direction $\bar{\epsilon}_z$ and where

$$c_2 = (b + \rho_b)^2 - c_1 b^2 \tag{4}$$

In these expressions for the strain components, r is the radius measured in the undeformed tube, b is the original outside radius, and ρ_b is the radial displacement of the outside of the tube.

When the tubular tests were interrupted for measurement of the axial strain and the increase in diameter, these measured values were used to calculate the distribution of the strain components. In Fig. 2 the strain distribution in a typical specimen after 358 hours is shown. For this example the axial strain $\bar{\epsilon}_z$ was 0.017 and the increase in outside diameter was 0.010 in. ($\rho_b = 0.005$). In this specimen, which is typical of several so examined, the axial strain and increase in diameter was rather low. These low values were chosen deliberately to try to show the correct strain distribution during second stage creep. If readings had been used from specimens that were near the rupture point, the average values of axial strain or diameter measurements would not have given representative results.

Another reason for making the analysis while the strains are small is the change that takes place in the stress ratios as deformation proceeds. The tangential stress increases more rapidly than the axial stress as the inside diameter increases during the strain.

As shown in Fig. 2, the radial and tangential strains are much higher at the inside radius than at the outside surface. The effective strain which is defined as

$$\bar{\epsilon}_{eff} = \frac{\sqrt{2}}{3} \left[(\bar{\epsilon}_\theta - \bar{\epsilon}_r)^2 + (\bar{\epsilon}_r - \bar{\epsilon}_z)^2 + (\bar{\epsilon}_z - \bar{\epsilon}_\theta)^2 \right]^{1/2} \tag{5}$$

so that $\bar{\epsilon}_{eff}$ reduces to $\bar{\epsilon}_z$ for pure tension also varies with the radius and is greatest at the inside surface $r = a$. Under steady-state conditions such as are encountered during the second stage of creep the effective strain rate will vary with the radius in the same manner as the effective strain.

In a thin-walled tubular specimen it is quite logical to use average values of the radial, axial, and tangential stresses to calculate an effective stress. It is quite logical to use the strains at the mean diameter to calculate an effective strain or effective strain rate. In the present analysis we shall develop expressions for the stress components and the effective strain rate as functions of the radius in order to compare the tubular tests with the uniaxial tension tests.

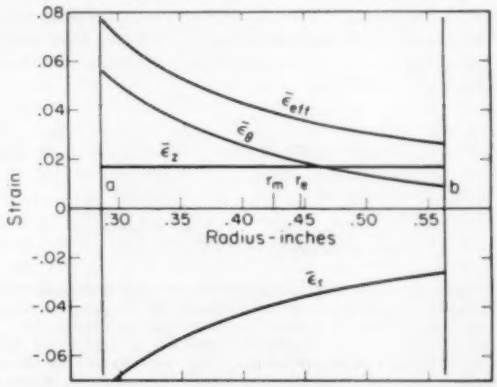


Fig. 2 Strain distribution in a typical specimen after 358 hours at load

For the case of a thin-walled tube under equal axial and tangential stresses the radial strain should be negative and equal to twice the axial or the tangential strain. It would be logical to compare these strains at the mean radius. In Fig. 2 it will be noted that at the mean radius r_m the tangential strain $\bar{\epsilon}_\theta$ is greater than the axial strain $\bar{\epsilon}_z$, and it appears that some other radius should be chosen as an equivalent radius. Since the effective strain [Equation (5)] includes the three principal strains in its make-up, it might be more logical to choose that radius which divides the cylinder into equal volumes. This radius we will call the equivalent radius r_e.

$$r_e = \left[\frac{b^2 + a^2}{2}\right]^{1/2} \tag{6}$$

We will hope, then, that this equivalent radius will be the radius at which the axial and tangential strains are equal for the case where the average axial and average tangential stresses are equal. This choice of equivalent radius does not have much theoretical background and is only suggested because some such assumption must be made in order to proceed with the analysis of the stress distribution. The necessity of this assumption will become evident later when the stress distribution is investigated.

A corresponding assumption must be made for the case of the pure internal pressure tests. In a thin-walled tube under pure internal pressure it can be shown theoretically and experimentally that there is no extension in the axial direction. In the thick-walled tube this is not so obvious but it might be assumed as a beginning for a trial and error solution. The calculated and measured resultant end loads could then be compared as a boundary condition.

The analysis so far has been in terms of natural strain. It has been found, however, that in order to determine the stress distributions it is too difficult to proceed with the more acceptable strains and that for convenience sake it is advisable to use conventional strains. It will be obvious later when the actual creep curves are presented that the strains are rather small in the second stage creep range, and therefore the analysis based on conventional strains will be sufficiently accurate.

In order to determine the stress distribution from the measured and computed strain rate distribution it is necessary to adopt some governing creep law. Such a law might be expressed in terms of the effective stress σ_{eff} and effective strain rate $u = \partial\epsilon_{eff}/\partial t$. These variables are defined by the relationships

$$\sigma_{eff} = \frac{1}{\sqrt{2}}\left[(\sigma_\theta - \sigma_r)^2 + (\sigma_r - \sigma_z)^2 + (\sigma_z - \sigma_\theta)^2\right]^{1/2} \tag{7}$$

$$u_{eff} = \frac{\sqrt{2}}{3}\left[(u_\theta - u_r)^2 + (u_r - u_z)^2 + (u_z - u_\theta)^2\right]^{1/2} \tag{8}$$

wher the stresses are based on the original dimensions and the strain rates are the time derivatives of the conventional strains.

Either a power function relationship or a hyperbolic sine relationship might be used. In this analysis the power function has been used because it leads to an expression that can be integrated without difficulty for the case of zero axial extension.

$$u_{eff} = \frac{\partial\epsilon_{eff}}{\partial t} = B\sigma_{eff}^n \tag{9}$$

The relationships between the components of the strain rate and the stress components are assumed to be

$$u_\theta = \frac{1}{\psi}\left[\sigma_\theta - \frac{\sigma_r + \sigma_z}{2}\right] \tag{10}$$

$$u_r = \frac{1}{\psi}\left[\sigma_r - \frac{\sigma_z + \sigma_\theta}{2}\right] \tag{11}$$

$$u_z = \frac{1}{\psi}\left[\sigma_z - \frac{\sigma_\theta + \sigma_r}{2}\right] \tag{12}$$

where ψ is some function of the stress and hence some function of the radius. From Equation (12) the axial stress is

$$\sigma_z = \psi u_z + \frac{\sigma_r + \sigma_\theta}{2} \tag{13}$$

and when this is substituted in Equation (7) the effective stress becomes

$$\sigma_{eff} = \frac{1}{2}\left[3(\sigma_\theta - \sigma_r)^2 + 4\psi^2 u_z^2\right]^{1/2} \tag{14}$$

Since the sum of the three strain rate components is zero, the effective strain rate can be written as

$$u_{eff} = \frac{2}{\sqrt{3}}\left[u_\theta^2 + u_\theta u_z + u_z^2\right]^{1/2} \tag{15}$$

The main reason for reverting to the conventional strains is to be able to use the compatibility equation for small strains

$$r\frac{\partial\epsilon_\theta}{\partial r} + \epsilon_\theta - \epsilon_r = 0 \tag{16}$$

The use of this equation to determine the corresponding expression for the strain rates will lead to a more simplified expression for the effective strain rate than would be obtained from Equations (1), (2), and (3).

The compatibility equation for the strain rates is obtained from that of the strains by differentiating it term by term

$$r\frac{\partial u_\theta}{\partial r} + u_\theta - u_r = 0 \tag{17}$$

Since u_r can be expressed in terms of u_θ and u_z, which does not depend upon r, Equation (17) can be integrated. The expressions for the strain rates then become

$$u_\theta + \frac{u_z}{2} = \frac{c_2}{r^2} \tag{18}$$

$$u_r + \frac{u_z}{2} = -\frac{c_1}{r^2} \tag{19}$$

The effective strain rate can now be expressed in terms of the radius and the axial strain rate.

$$u_{eff} = \frac{1}{\sqrt{3}\,r^2}\left[4c_2^2 + 3u_z^2 r^4\right]^{1/2} \tag{20}$$

Equations (10) and (12) can be combined with Equation (18) to produce a relationship between the stress difference $(\sigma_\theta - \sigma_r)$ and the radius r.

$$u_\theta + \frac{u_z}{2} = \frac{3}{4\psi}(\sigma_\theta - \sigma_r) = \frac{c_2}{r^2} \tag{21}$$

from which

$$\psi = \frac{3(\sigma_\theta - \sigma_r)r^2}{4c_2} \tag{22}$$

and when this is substituted in Equation (14) the effective stress becomes

$$\sigma_{eff} = \frac{\sqrt{3}\,(\sigma_\theta - \sigma_r)}{4c_2}\left[4c_2^2 + 3u_z^2 r^4\right]^{1/2} \tag{23}$$

When Equations (20) and (23) are now substituted in the creep law, Equation (9), we obtain a final expression for the stress difference $(\sigma_\theta - \sigma_r)$.

$$(\sigma_\theta - \sigma_r) = \frac{4c_3}{B^{1/n}(3)^{\frac{n+1}{2n}} r^{2/n}(4c_3^2 + 3u_s^2 r^4)^{\frac{n-1}{2n}}} \tag{24}$$

The equilibrium equation for radial symmetry is

$$r\frac{\partial \sigma_r}{\partial r} + \sigma_r - \sigma_\theta = 0 \tag{25}$$

The radial stress σ_r can be found by substituting Equation (24) in the expression

$$\sigma_r = \int_b^r \frac{(\sigma_\theta - \sigma_r)}{r}\, dr \tag{26}$$

Equations (26) and (24) are used to determine the radial and the tangential stresses. The axial stress can then be determined from Equations (22) and (13).

It will be noted that these expressions contain the axial strain rate which is known for the biaxial tests and is assumed zero for the pure internal pressure test. The constant c_3 can be evaluated if some relationship between u_θ and u_s is assumed and if u_θ at some given value of r is known. The procedure to be followed, then, depends upon the type of loading. For the pure internal pressure tests it is assumed that u_s is zero and that u_θ at $r = b$ can be obtained from the creep curves. For the biaxial tests it is assumed that $u_\theta = u_s$ at some given value of the radius (say $r = r_s$). The constant c_3 can then be determined in terms of u_s which can be determined from the creep curves.

Creep Tests in Uniaxial Tension

The data from the uniaxial tension tests are listed in Table 2. The minimum creep rates are plotted in Fig. 3. The exponent for the power function relationship of Equation (9) is $n = 6.29$. This particular lot of type 316 stainless steel was rather soft. In each specimen plastic deformation occurred on loading before the stress reached 12000 psi.

Only 3 of the specimens fractured. There was very little localized necking, but many cracks opened in the vicinity of the fracture. This accounts for the low values of reduction in area in otherwise ductile material. It is interesting to note that at a stress of 18300 psi the transition point occurred at 4 per cent strain and at 220 hours, while at a stress of 11450 there is no indication of third stage creep after 10200 hours and 4.9 per cent strain.

Creep Tests on Tubular Specimens

Of the 10 tubular specimens tested, 5 were loaded with pure internal pressure and 5 with equal biaxial tensions. In the biaxial tension tests the average axial stress was equal to the average tangential stress. The creep curves for the biaxial tension tests are shown in Fig. 4 while those for the pure internal pressure tests are shown in Figs. 5 and 6. Specimens T-8, T-9, and T-10 are extra heavy-walled tubes. The extensometers could not be left on the specimens until fracture occurred, hence the rupture points are not generally shown on the creep curves. The data relating to these tests are shown in Table 3.

Since there is localized flow in both the axial and the tangential directions in the neighborhood of the fracture, the strain at this point is not easily determined. The fracture strains listed in this report have been obtained in the following manner. After fracture the wall thickness at the edge of the rupture was measured. From these measurements an average conventional radial strain was calculated, and from this conventional strain the corresponding natural strain was determined. For the tests under biaxial tension it was assumed that the strain in the axial direction was equal to the strain in the tangential direction and that each was equal to one half the average radial strain but of opposite sign. The average radial strain then corresponds to the effective strain at rupture. For the pure internal pressure test it was assumed that the axial strain was zero. The tangential strain at fracture then is equal to the radial strain but of the opposite sign. These average values were then substituted in Equation (5) to obtain the effective strain at fracture. The values of effective strain at rupture listed in Table 3 appear to be erratic, and this, no doubt, is due to the large number of cracks in the vicinity of the fracture. Some idea of the cracking near the fracture may be gained from the photograph of a ruptured specimen shown in Fig. 7.

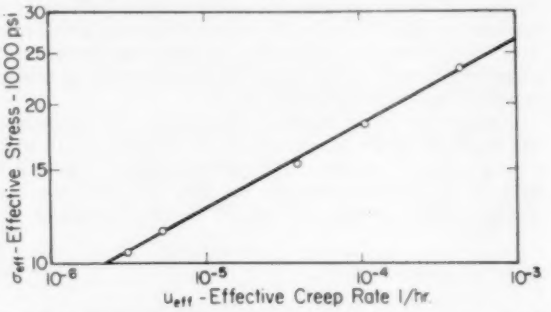


Fig. 3 Stress versus creep rate for uniaxial tension tests on type 316 stainless steel at 1200 F

Table 2 Test data for uniaxial tension creep test

Type 316 stainless steel

Specimen no.	1	2	3	4	5	6
Temperature, deg F	1200	1200	1200	1200	1200	1200
Stress, psi	23400	18300	28300	15300	10450	11450
Rupture, hr	213	1248	37.5	>620	>868	>10200
Rupture strain, per cent	28.0	39.3	21.5	...	...	...
Minimum creep rate, 1/hr	4.4×10^{-4}	1.1×10^{-4}	2.8×10^{-3}	3.9×10^{-5}	3.1×10^{-6}	5.2×10^{-6}
Reduction of area	29.7[b]	31.9[b]	23.8[b]	...	...	...
Time to 0.5% strain, hr	0	0	0	0	1000[c]	800
Time to 1.0% strain, hr	0	0	0	125	...	1800
Time to 3.0% strain, hr	5	145	1	625[c]	...	5820
Transition point, hr	55	220	15	...	...	...
Transition strain, per cent	6	4	9	...	...	>4.9
Original Vickers hardness	146	139	139	140	150	155
Effective strain at rupture, per cent[a]	35.2	38.5	27.2	...	...	...

[a] Natural strain, based on reduction in area.
[b] Reduction in area figures are low due to the large number of voids opened within the material.
[c] Extrapolated.

Table 3 Test data for creep tests on tubular specimens

Type 316 stainless steel

Specimen no.	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9	T-10
Temperature, deg F	1200	1200	1200	1200	1200	1200	1200	1200	1200	1200
Pressure, psi	15000	14000	13400	16000	13300	14600	12000	24000	23000	21000
Aver. axial stress, psi	5040	4660	4480	16100	13350	14700	12000	3020	11540	2640
Aver. tangential stress, psi	15100	14000	13400	16100	13350	14700	12000	12030	11540	10520
Aver. radial stress, $= -p/2$	−7500	−7000	−6700	−8000	−6650	−7300	−6000	−12000	−11500	−10500
Rupture time, hr	400	970	1590	225	1018	470	>1828	639	680	1550
Transition point, hr	60	0	0	75	140	90	250	100	105	0
Orig. wall thickness, in.	0.284	0.287	0.303	0.281	0.282	0.281	0.282	0.375	0.375	0.375
Wall thickness at fracture, in.	0.172	0.222	0.221	0.201	0.223	0.187	...	0.217	0.247	...
Aver. radial strain at rupture[a]	−50.0	−25.8	−31.6	−33.6	−23.4	−40.8	...	−54.6	−41.8	...
Effective strain at rupture[a]	57.7	29.8	36.5	33.6	23.4	40.8	...	63.0	41.8	...
Axial strain rate, u_z[b]	...	...	...	1.2×10^{-4}	2.0×10^{-5}	5.4×10^{-6}	6.5×10^{-6}	...	3.3×10^{-6}	...
Diametral strain rate, $\frac{\partial \rho_0}{\partial l}$[b]	1.74×10^{-5}	7.3×10^{-6}	5.0×10^{-6}	...	...	...	...	1.07×10^{-6}	...	5.0×10^{-6}
Effective aver. stress, $(\sigma_{av})_{eff}$, psi	19600	18200	17500	24100	20000	22000	18000	21000	23100	18250
Equivalent radius, r_e, in.	0.450	0.455	0.478	0.446	0.446	0.445	0.446	0.420	0.419	0.420
Effective strain rate at r_e[b]	6.2×10^{-5}	2.4×10^{-5}	1.8×10^{-5}	2.5×10^{-4}	4.0×10^{-5}	1.1×10^{-4}	1.25×10^{-5}	4.0×10^{-5}	6.5×10^{-5}	1.9×10^{-5}

[a] Natural strains.

[b] Minimum creep rates, 1/hr.

Evaluation of Stress Components

The stress distribution in the tubular specimens was calculated according to the equations developed earlier in this paper. For the pure internal pressure tests, u_z in Equation (24) was assumed to be zero which made it possible to integrate Equation (26) analytically. For the biaxial tension tests Equation (26) was integrated numerically. The summation was made by starting at $r = b$ where σ_r was known to be zero.

The first step in evaluating the stresses was the evaluation of the constant n in Equation (9) from the uniaxial tension tests. The value is obtained from the slope of the curves in Fig. 3. There may be some arguments in favor of determining the value of both B and n from the uniaxial tension tests. If both constants were determined from Fig. 3, the calculated points for the biaxial tests would fall directly on the line drawn in Fig. 3, and any discrepancy between test results and theory would show up as a difference between a calculated and a measured internal pressure or axial load. Since both the effective stress and the effective strain rates vary through the wall of the tubular specimen, the results of a test on such a specimen will show up as a segment of a line or curve rather than as a point in a plot like that of Fig. 3. In the present analysis the exponent n was determined from the uniaxial tension tests while the constant B was allowed to take on the value that would make the calculated end load equal to the measured axial load. The result of each tubular test will then show up as a straight line segment parallel to the curve for the uniaxial tension tests in a plot like that of Fig. 3.

The calculation of the stresses in a biaxial tension test can then be carried out in the following manner. The constant c_3 is evaluated in Equation (18) by assuming that u_θ is equal to the known u_z at $r = r_s$. The effective strain rate u_{eff} can then be obtained as a function of the radius from Equation (20). Since the value of the constant B in Equation (9) is not to be evaluated from the uniaxial tension tests, and since by Equation (18) the constant c_3 is proportional to the axial rate u_z, all of the constants in Equation (24) including the axial rate u_z can be grouped in one unknown constant A and this equation can be written as

$$\frac{(\sigma_\theta - \sigma_r)}{A} = f(r) \tag{27}$$

where $f(r)$ is a known function of the radius. Integration of Equation (26) then yields the value of σ_r/A and the constant A

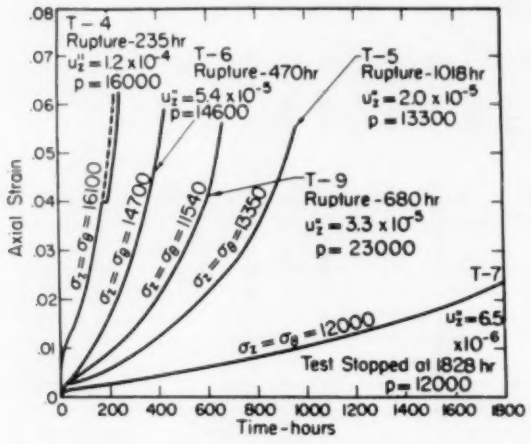


Fig. 4 Creep of type 316 stainless steel at 1200 F, biaxial tension

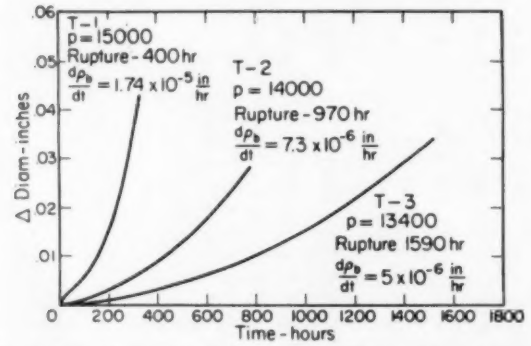


Fig. 5 Creep of type 316 stainless steel at 1200 F, pure internal pressure

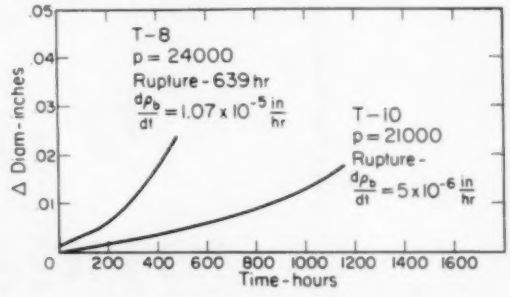


Fig. 6 Creep of type 316 stainless steel at 1200 F, pure internal pressure

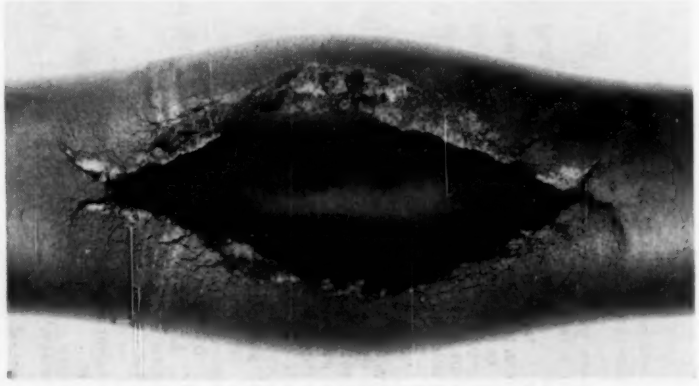


Fig. 7 Typical fracture of specimen under biaxial tensions

can be evaluated by setting σ_r equal to $-p$ at $r = a$. The tangential stress σ_θ is then obtained from Equation (24), and the axial stress σ_z from Equations (13) and (22). The effective stress σ_{eff} can then be evaluated from Equation (23).

If the calculated end load obtained from integrating the axial stress over the cross-sectional area agrees with the load determined by the applied tensile load and the internal pressure, the stress distribution is assumed to be correct. If the two loads do not agree, a new estimate of the radius at which the tangential and the radial strain rates are equal must be made and the work repeated until agreement between the two loads is reached. A comparison of the measured and calculated loads is shown in Table 4. Where two listings are shown for one specimen, the first set of values represent the original estimate of r^* the radius at which it was assumed that u_θ was equal to u_z and the second represents an improved estimate of r^*.

Table 4 Comparison of measured and calculated axial loads

Specimen	Pressure	External load	Measured load	Calculated load	r_a	r^*
316 T-1	15000	0	3840	3820	...	...
316 T-2	14000	0	3640	3645	...	...
316 T-3	13400	0	3840	3730	...	...
316 T-4	16000	7950	11950	12600	0.446	0.446
316 T-4	16000	7950	11950	11600	0.446	0.455
316 T-5	13300	6650	9975	9850	0.446	0.446
316 T-6	14600	7270	10920	10900	0.445	0.445
316 T-7	12000	6000	9000	8950	0.446	0.446
316 T-8	24000	0	2660	2685	...	...
316 T-9	23000	7650	10200	10170	0.419	0.419
316 T-10	21000	0	2330	2430	...	...

In this table r^* is the radius at which it was assumed that u_θ was equal to u_z.

The agreement of the calculated and measured values of the axial loads for the pure internal pressure tests was fairly good as can be seen in Table 4. These values, however, could be made to check more closely by assuming a value other than zero for the axial strain. This has not been done, however, for no measurable amount of axial strain was observed in any of these specimens.

Typical stress distribution curves for a biaxial tension test and a pure internal pressure test are shown in Figs. 8 and 9. These curves are for the specimens with the heaviest walls and the highest pressures. For the pure internal pressure test the three stress components gradually increase with the radius. The maximum tensile stress is the tangential stress at the outside radius. For the biaxial tension tests the axial stress was less than the tangential stress at the inside surface but increased more rapidly and became greater than the tangential stress at the outside surface where it was the greatest tensile stress in the tube.

Comparison of Uniaxial and Biaxial Results

The minimum creep rates for the uniaxial tests shown in Fig. 3 are replotted in Figs. 10 and 11 and the effective stresses and strain rates for the tubular specimens have been plotted there also.

The data are plotted on two curve sheets for ease in reading. The tubular test results show lower strain rates than would be predicted from the uniaxial tension tests. This discrepancy appears to become greater as the stress is reduced. As far as minimum creep rate is concerned, there seems to be no significant difference between the results of the pure internal pressure and the biaxial tension tests. For example, in Fig. 11, the specimens which represent the best and the worst agreement with the uniaxial tension tests were both loaded by pure internal pressure.

The manner in which the results of the tubular specimens deviate from the results of the uniaxial tension tests is not fully understood and no adequate explanation can be given at this time. If the tubular specimens were to be analyzed without the uniaxial tension tests, it appears that a better fit would be obtained by using a larger value of n.

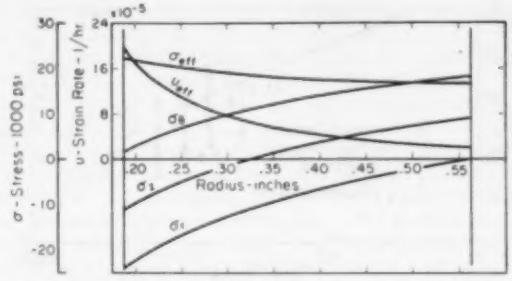


Fig. 8 Stress distribution in thick-walled specimen T-8, type 316 stainless steel at 1200 F, pure internal pressure

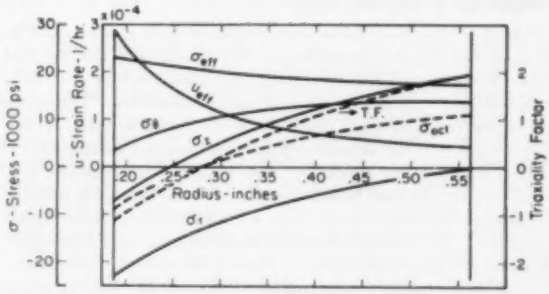


Fig. 9 Stress distribution in thick-walled specimen T-9, type 316 stainless steel at 1200 F, biaxial tension

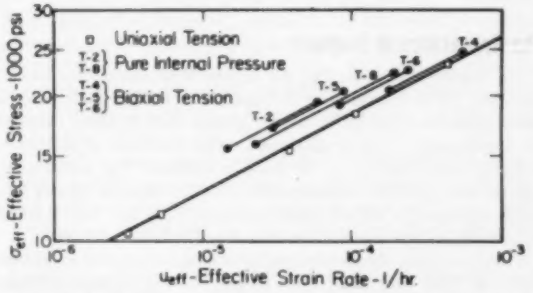

Uniaxial Tension
T-2, T-8 Pure Internal Pressure
T-4, T-5, T-6 Biaxial Tension
σeff - Effective Stress - 1000 psi
ueff - Effective Strain Rate - 1/hr.

Fig. 10 Comparison of uniaxial and biaxial creep tests type 316 stainless steel at 1200 F

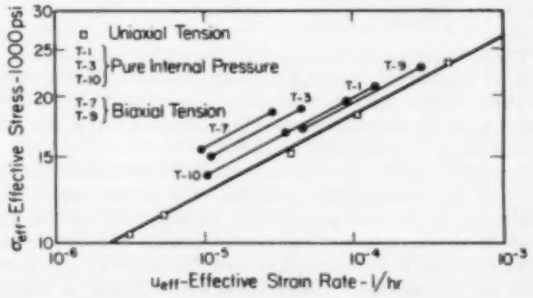

Uniaxial Tension
T-1, T-3, T-10 Pure Internal Pressure
T-7, T-9 Biaxial Tension
σeff - Effective Stress - 1000 psi
ueff - Effective Strain Rate - 1/hr

Fig. 11 Comparison of uniaxial and biaxial creep tests type 316 stainless steel at 1200 F

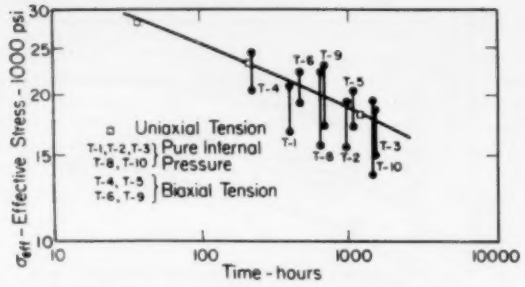


Fig. 12 Comparison of rupture times, type 316 stainless steel at 1200 F

Comparison of Rupture Times

A plot of effective stress versus rupture time can be used to compare the rupture properties of the biaxial and the uniaxial tests. Such a curve is shown in Fig. 12. Each rupture of a tubular specimen is represented by a straight vertical line. The upper end of the line represents the stress at the inside surface. If the effective stress at the outside surface were used as a means of comparison, the tubular specimens would have a low rupture life. If the comparison were based upon the stress at the inside surface, the tubular specimens would have a longer life than the uniaxial specimens. The stress at the equivalent radius is not much higher than it is at the outside surface. Thus if the stress at the equivalent radius were used for comparison purposes, the tubular specimens would still have an unfavorable life. This shorter life for the tubular specimens is partly due to the large increase in tangential stress during the third stage of creep.

The Appearance of Fracture

The fracture shown in Fig. 7 is typical of all the biaxial tension fractures. The fractures in the specimens loaded with pure internal pressure were quite similar except that in those cases all the small cracks were longitudinal. The fractures all started as grain boundary cracks at the outside surface. The oxide coatings on the fracture surfaces show that the surface cracks that ultimately lead to failure progress about $^2/_3$ of the way through the wall at a slow rate before catastrophic failure occurs.

Examination of the outside surfaces of the fractured specimens indicates that practically the entire outside surface contains small grain boundary cracks. The inside surfaces also contained some longitudinal cracks, but these were not as deep as the external cracks and they may not have formed until the time of the actual fracture when the tube wall is subjected to rather violent bending.

The fact that the destructive cracks started at the outer surface may be explained by the higher tensile stresses appearing there as shown in the stress distribution curves. This higher tensile stress seems to overshadow the higher effective stress and the greater strain that occurs at the bore.

It is quite possible that the atmosphere or environment had an effect on the grain boundary cracking. The outside surface was exposed to an air atmosphere, while the inside surface was exposed to an atmosphere of superheated steam. It should be remembered that there was no flow past this inside surface, however, and the atmosphere there would not be typical of that in actual steam piping. Apparently, though, the steam atmosphere that was present was not particularly damaging.

Another factor that may tend to cause the cracking to be more pronounced at the outer surface is the manner in which the hydrostatic component of the stress varies with the radius. The hydrostatic component of the stress is the normal stress on the octahedral planes σ_{oct} and is given by the expression

$$\sigma_{oct} = \frac{\sigma_1 + \sigma_2 + \sigma_3}{3} = \frac{\sigma_\theta + \sigma_r + \sigma_z}{3} \tag{28}$$

This quantity has been evaluated for one of the extremely heavy-wall specimens, T-9. The results are plotted in Fig. 9. The hydrostatic stress is 8800 psi compression at the bore and 11000 psi tension at the outside surface. It is known that the ductility of a material is reduced if the ratio of maximum tensile to maximum shearing stress is increased. A triaxiality factor[1] which would be analogous to this ratio and which would contain the stresses in the three directions could be written as

$$\text{TF} = \frac{\sqrt{2}\,\sigma_{oct}}{\tau_{oct}} = \frac{\sqrt{2}\,(\sigma_1 + \sigma_2 + \sigma_3)}{[(\sigma_1 - \sigma_3)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2]^{1/2}} \tag{29}$$

The coefficient in this ratio is introduced to make the triaxiality factor equal to $+1$ for uniaxial tension. It then has the value of zero for torsion and -1 for uniaxial compression. In Fig. 9 this triaxiality factor is also plotted against the radius. It varies from a value of -1.15 at the bore to a value of $+1.925$ at the outside. Thus it would be expected that the material at the outside surface would be less ductile than that near the inner surface.

Practical Use of Test Data

Since the determination of the test data shown in Figs. 10, 11, and 12 required the integration of Equation (26), and since this in some cases had to be done numerically, it would be convenient if the data could be interpreted in a more simplified manner. This might be done if an effective stress based upon the average values of the three principal stress components would be used. Such an effective average stress could be defined as

$$\sigma_{(av)eff} = \frac{1}{\sqrt{2}}\,[(\sigma_{\theta(av)} - \sigma_{r(av)})^2 + (\sigma_{r(av)} - \sigma_{z(av)})^2 + (\sigma_{z(av)} - \sigma_{\theta(av)})^2]^{1/2} \tag{30}$$

Where the average values of the stress components for the tubular specimens are determined from the following simple relationships

$$\sigma_{\theta(av)} = p\,\frac{a}{b - a} \tag{31}$$

$$\sigma_{r(av)} = -\frac{p}{2} \tag{32}$$

$$\sigma_{z(av)} = \frac{pa^2}{b^2 - a^2} + \frac{W}{\pi(b^2 - a^2)} \tag{33}$$

In these expressions W is the applied axial load, p is the internal pressure, a is the inside radius, and b is the outside radius. The value of this effective average stress has been calculated and is listed in Table 3. In order for such a stress to be useful, it is necessary to relate it to some particular value of the effective strain rate. For comparison with the behavior under uniaxial tension the effective average stress $\sigma_{(av)eff}$ is plotted against the effective strain rate at $r = r_e$ in Fig. 13. This type of plotting moves the test points for the tubular tests to the left or in the direction of more conservative design. For a given calculated value of the effective average stress the actual effective strain rate would be less than that calculated from uniaxial tension data.

[1] This factor has been introduced in a paper, "Stress Distribution and Plastic Deformation in Rotating Cylinders of Strain Hardening Material," by E. A. Davis and F. M. Connelly, *Journal of Applied Mechanics*, vol. 26, TRANS. ASME, vol. 81, 1959, pp 25–30.

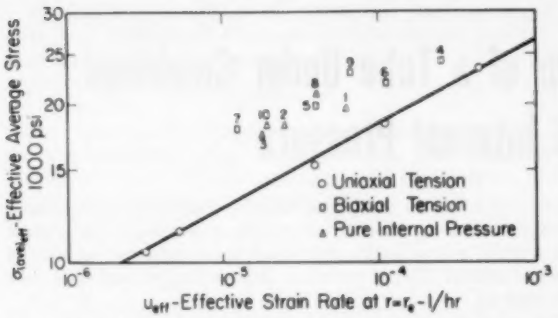


Fig. 13 Plot of effective average stress versus minimum creep rate at the equivalent radius for type 316 stainless steel at 1200 F

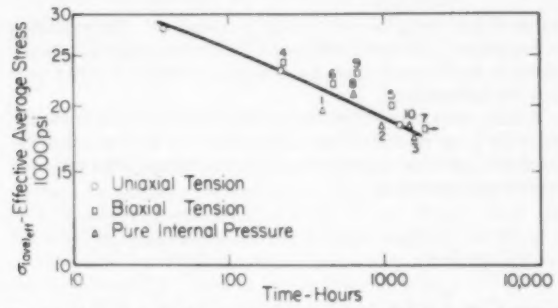


Fig. 14 Comparison of rupture times, type 316 stainless steel at 1200 F

This effective average stress can also be used in a time to rupture plot as shown in Fig. 14. In this plot the rupture points for all the biaxial tension tests fall on the safe side of the curve for uniaxial tension. It is of interest to note that the points for the pure internal pressure tests on the extra heavy-walled tubes also fall on the safe side of the curve, but that the pure internal pressure tests on tubes with OD to ID ratio of 2 have shorter rupture times than would be predicted by the uniaxial tension tests. It should be pointed out, however, that the difference between the pure internal pressure test and the uniaxial tension tests decreases as the stress decreases. Thus it would appear that the uniaxial tension creep rupture data are safe bases for design if the design stresses are calculated in terms of an effective average stress. This is particularly true, according to the present test data, if the design is to be based on creep rate criteria. It may not be completely true if rupture time is the basis for design, although it tends to come nearer to the true values of rupture time at smaller stresses.

It might be of interest here to compare the effective average stress approach to the design of steam piping with the formula given in the ASME Boiler and Pressure Vessel Code. The expression for the minimum wall thickness of steam piping as given in Paragraph P-23 of Section 1 on Power Boilers is

$$t_m = \frac{pD}{2SE + 2yp} + c \tag{34}$$

where t_m is the minimum wall thickness in inches, p is the maximum internal pressure in pounds per square inch, D is the outside diameter of the pipe, E is the efficiency of a longitudinal weld and will be considered here as unity, c is an allowance for corrosion and will be neglected in this comparison, y is a factor depending on the material and the temperature and for the present case is 0.7, and S is an allowable stress due to internal pressure. The value of S for type 316 stainless steel at 1200 F according to the Boiler Code is 6800 psi.

The maximum pressure for a tube whose outside diameter is twice the inside diameter can be determined from Equation (34) by letting $D = 4t$ and solving for p with S and y as prescribed. The pressure so determined is 5225 psi. The effective average stress for such a tube as obtained from Equations (30)–(33) is 6800 psi which agrees exactly with the value of S given by the boiler code.

According to the present test data and method of an lysis such a tube would creep at a rate lower than that of a uniaxial tension specimen at 6800 psi, but it might rupture in a slightly shorter time.

For the extra heavy-walled tubes where the outside diameter is 3 times the inside diameter the maximum allowable pressure as given by the Boiler Code formula is 8500 p i. The corresponding effective average stress for this tube is 7400 psi which is considerably higher than the allowable value of 6800 psi. When compared with uniaxial data the effective average stress would then lead to a much more conservative design than would be obtained from the Boiler Code formula. The test data indicate that a tubular specimen of such cross section would creep more slowly and rupture later than a uniaxial tension specimen at 7400 psi. There may be some question whether or not it would behave better than a tension specimen at 6800 psi.

Conclusions and Remarks

The analysis of the combined stress creep results presented here has shown that both combined stress and uniaxial stress results can be fitted to the same creep law. The analysis, however, did indicate a deviation of the combined stress results away from those for uniaxial tension. It is hoped that this deviation can be explained in the future.

It should be remembered that the analysis presented here is not the only possible or even the most preferable analysis. As mentioned earlier, the power function law was adopted because it led to expressions that could be integrated under favorable conditions.

A better analysis might be made if an extensometer capable of measuring both the axial strain and the increase in diameter could be developed.

In the present investigation the distribution of the strains could be completely determined by making use of certain theoretical considerations. The general problem of strain distribution thus was not studied. The assumptions involved in adopting Equations (10) to (12) were less important in these tests than would have been the case if other ratios of axial to tangential load had been tested.

The effective average stress as defined in Equations (30)–(33) provides a simple means of evaluating a state of combined stress in terms of uniaxial tension data and compares quite closely with the Boiler Code formula for tubes in which the outside diameter is twice the inside diameter.

For the tubes with the heaviest wall, OD = 3ID, the effective average stress leads to a more conservative design than that based on the Boiler Code formulas.

IAIN FINNIE
Engineer, Shell Development Company, Emeryville, Calif. Assoc. Mem. ASME

Steady Creep of a Tube Under Combined Bending and Internal Pressure

It is shown that for a tube in bending the steady-state bending creep rate may be greatly increased by the addition of internal pressure. If the maximum bending stress is small compared to the pressure stress, a very simple expression is obtained for the ratio of bending creep rates with and without internal pressure. Experiments on lead tubes show good agreement with the theory.

FOR MANY MATERIALS at elevated temperatures the curve of strain, as a function of time in a tension test at constant temperature, shows a well-defined region of steady-state creep. It is then often permissible for convenience in strain calculations to replace the strain-time curve by the straight line shown in Fig. 1. The creep rate is then only a function of stress. The method of predicting strains under multiaxial stresses from these tensile data is fairly well established. Solutions have been given, for example, for thin and thick-walled cylinders under internal pressure [1],[1] tubes in bending [2], and rotating disks [3].

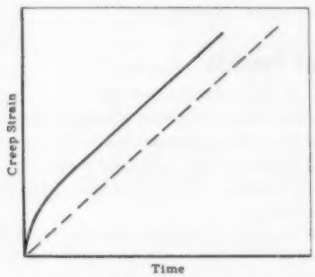


Fig. 1 Schematic constant stress creep curve and straight line idealization

An interesting case of combined stress creep occurs when bending moments and internal pressure act together on a thin-walled tube. This problem was originally mentioned by Bailey [4] and his comprehensive 1931 creep paper, but has not received subsequent attention in the literature. We will show that the bending strain rate under the combined loading may be considerably greater than that due to bending alone. Such a magnification, which of course does not occur with elastic stresses, may be of considerable importance in the design of high-temperature piping.

Analysis of Strains

As a starting point we assume that the steady-state creep rate $\dot{\epsilon}$ and stress σ in a tensile test are related by the empirical creep law

$$\dot{\epsilon} = B\sigma^n \tag{1}$$

where B and n may be functions of temperature. The extension of Equation (1) to multiaxial strain prediction is given in many places in the literature (see, for example, reference [1]) and need only be outlined here.

Briefly, one assumes that the stresses and strain rates in steady-state creep are related by the expression $\dot{\epsilon}^* = B\sigma^{*n}$ in which the invariant quantities $\dot{\epsilon}^*$ and σ^* are given in terms of the principal stresses and strains by

$$\sigma^* = \frac{1}{\sqrt{2}}\left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2\right]^{1/2}$$

$$\dot{\epsilon}^* = \frac{\sqrt{2}}{3}\left[(\dot{\epsilon}_1 - \dot{\epsilon}_2)^2 + (\dot{\epsilon}_2 - \dot{\epsilon}_3)^2 + (\dot{\epsilon}_3 - \dot{\epsilon}_1)^2\right]^{1/2}$$

Making the additional conventional assumptions that the material is incompressible during plastic deformation (creep) and that at any location the principal shear strain rates are proportional to and in the same direction as the principal shear stresses leads to two further equations

$$\dot{\epsilon}_1 + \dot{\epsilon}_2 + \dot{\epsilon}_3 = 0$$

and

$$\frac{\dot{\epsilon}_1 - \dot{\epsilon}_2}{\sigma_1 - \sigma_2} = \frac{\dot{\epsilon}_2 - \dot{\epsilon}_3}{\sigma_2 - \sigma_3} = \frac{\dot{\epsilon}_3 - \dot{\epsilon}_1}{\sigma_3 - \sigma_1} = C$$

Combining these two equations leads to

$$\dot{\epsilon}_1 = \frac{2}{3}C\left[\sigma_1 - \frac{1}{2}(\sigma_2 + \sigma_3)\right]$$

and two similar expressions for $\dot{\epsilon}_2$ and $\dot{\epsilon}_3$. Using the assumed relationship between stress and strain-rate invariants, this may be written

$$\dot{\epsilon}_1 = B\sigma^{*n-1}\left[\sigma_1 - \frac{1}{2}(\sigma_2 + \sigma_3)\right]$$

For a tube under combined bending and internal pressure the principal stresses, σ_t, σ_a, σ_r, lie in the tangential, axial, and radial directions. If we consider a thin-walled tube, and thus ignore σ_r, the axial strain rate $\dot{\epsilon}_a$ at a given location may therefore be written as

$$\dot{\epsilon}_a = B(\sigma_t^2 - \sigma_a\sigma_t + \sigma_a^2)^{\frac{n-1}{2}}\left(\sigma_a - \frac{\sigma_t}{2}\right) \tag{2}$$

The tangential stress may be written directly in terms of the internal pressure p, inside tube radius r, and tube wall thickness t as $\sigma_t = pr/t$, but the distribution of the axial stress has yet to be determined.

[1] Numbers in brackets designate References at end of paper.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 3, 1959. Paper No. 59—Met-5.

We make the conventional assumption that plane sections remain plane in bending.[2] The variation of axial strain rate over the cross section can then be written as

$$\dot{\epsilon}_a = \dot{\epsilon}_{a1} y / y_1 \tag{3}$$

where $\dot{\epsilon}_a$ and $\dot{\epsilon}_{a1}$ are the axial strain rates at distances y and y_1 from the neutral axis for bending. Later we will take y_1 as the distance to the outer fiber. If the corresponding axial stresses are σ_a and σ_{a1}, Equations (2) and (3) combined lead to

$$[\sigma_t^2 - \sigma_a\sigma_t + \sigma_a^2]^{\frac{n-1}{2}} \left[\sigma_a - \frac{\sigma_t}{2}\right] = [\sigma_t^2 - \sigma_{a1}\sigma_t + \sigma_{a1}^2]^{\frac{n-1}{2}} \left[\sigma_{a1} - \frac{\sigma_t}{2}\right] \frac{y}{y_1} \tag{4}$$

We can simplify the form of Equation (4) by writing

$$\sigma_a = \frac{\sigma_t}{2} + b\sigma_t \qquad \sigma_{a1} = \frac{\sigma_t}{2} + b_1\sigma_t \tag{5}$$

Since the axial stress due to internal pressure acting alone is $\sigma_a = \sigma_t/2$, we can interpret $b\sigma_t$ as the amount by which the combined loading increases the axial stress due to pressure alone. It is *not* necessarily the stress that would be produced by the bending moment acting alone in the absence of internal pressure. Inserting Equation (5) in Equation (4) leads to

$$b(0.75 + b^2)^{\frac{n-1}{2}} = b_1(0.75 + b_1^2)^{\frac{n-1}{2}} \frac{y}{y_1} \tag{6}$$

Two other equations may be obtained by considering force and moment equilibrium. Although a solution of these equations and Equation (6) is then possible in principle, it is not convenient in practice.

A greatly simplified solution is possible for the very common practical condition in which the bending stresses are small compared to the tangential stress due to internal pressure. If this is the case, we can neglect b_1^2 and b^2 relative to 0.75 and Equation (6) becomes

$$b = b_1 y / y_1 \tag{7}$$

That is, the additional stress due to bending varies linearly over the cross section and hence may be calculated as if it were an elastic stress. On the other hand, if the bending stresses are large compared to pressure stresses, then the bending stress distribution approximates that due to bending acting alone. Robinson [2] has shown that the ratio of the maximum stress σ_c in creep bending of a circular cross section to the maximum elastic stress σ_e which would be produced by the same bending moment is

$$(\sigma_e/\sigma_c) = \frac{8}{\left(3 + \frac{1}{n}\right)\sqrt{\pi}} \left[\frac{1 - (R_i/R_o)^{3+1/n}}{1 - (R_i/R_o)^4}\right] \frac{\Gamma\left(1 + \frac{1}{2n}\right)}{\Gamma\left(1.5 + \frac{1}{2n}\right)} \tag{8}$$

where R_i and R_o are the inner and outer radii and Γ is the gamma function.

The axial strain rate at the outside fiber distance y_1 from the neutral axis under combined bending and pressure is given by Equations (2) and (5) as

$$\dot{\epsilon}_{a1} = B\sigma_t^n(0.75 + b_1^2)^{\frac{n-1}{2}} b_1$$

The strain rate $\dot{\epsilon}_{a1}$ at the outer fiber due to bending acting alone would be

$$\dot{\epsilon}_{a1}' = B\sigma_t^n b_e^n$$

where $b_e = \sigma_e/\sigma_t$ and σ_e may be obtained by combining Equation (8) and an elastic bending calculation.

We can now write simple expressions for the ratio of the axial strain rates with and without internal pressure for two cases: Maximum bending stress small relative to tangential pressure stress ($b_1^2 \ll 0.75$).

$$\dot{\epsilon}_{a1}/\dot{\epsilon}_{a1}' \approx \left(\frac{0.75}{b_e^2} + 1\right)^{\frac{n-1}{2}} \left(\frac{\sigma_e}{\sigma_c}\right)^n \tag{9}$$

where σ_e/σ_c is obtained from Equation (8) and $b_e = \sigma_e/\sigma_t$. Maximum bending stress large relative to tangential pressure stress ($b_1^2 \gg 0.75$).

$$\dot{\epsilon}_{a1}/\dot{\epsilon}_{a1}' \approx \left(1 + \frac{0.75}{b_c^2}\right)^{\frac{n-1}{2}} \tag{10}$$

Experiments

To study the use of the preceding equations, tubes were tested under combinations of internal pressure and bending at 120 F. The material used for the experiments was lead containing 6 per cent antimony. When tested in a constant temperature water bath at 120 F, this material was found to give very reproducible creep data. For the stresses studied (300 to 1000 psi) it exhibited a very well-defined steady-state creep between about 20 hours after loading and the end of the test at 100 hours. To decrease the possibility of metallurgical changes during the test the specimens were held at 160 F for several hours before being tested at 120 F. While lead is far from a high-temperature material, its use has many advantages. Its creep curves at low temperatures and stresses are reasonably similar to those of heat resistant alloys at high temperatures and stresses. Methods of multiaxial stress analysis based on tension data, if they are to be generally applicable, should apply to all creeping materials provided the tension creep curves are reasonably similar. Strain measurements during the test were made by wire resistance strain gages glued to the specimens. Comparing the strains of specimens with and without strain gages showed that the stiffening effect of the gage could be neglected.

For a time scale in hours and stresses in psi, the expression fitted to the tension test data was

$$\dot{\epsilon} = 8 \times 10^{-12}\sigma^{2.66} \tag{11}$$

It is perhaps worth pointing out that an expression such as Equation (11) may be obtained from a single bending test with strain gages at several locations rather than from a number of tension and/or compression tests.

Fig. 2 shows strain readings for three axial locations on a tube loaded first by bending moment alone and later by the same bending moment combined with internal pressure. Since most of the data points lie within the widths of the lines in Fig. 2, they are not shown. The maximum tension and compression creep rates at the three cross sections for the two loading conditions are shown in Table 1. Also shown in Table 1 are the calculated maximum tensile stresses for elastic bending and creep bending and the ratios of these stresses to the tangential pressure stress (i.e., b_e and b_c). In this particular test, the stresses were chosen so that the values of b would lie between those for which Equation (9)

[2] This is only strictly true for pure bending. For bending moments which do not vary too rapidly, we assume that the action of shear forces in distorting the otherwise plane section may be neglected.

Table 1 Axial creep rates and stresses corresponding to Fig. 1

Location	Creep rate 21–46 hr bending only, $hr^{-1} \times 10^6$	Creep rate 69–99 hr bending and pressure, $hr^{-1} \times 10^6$	Elastic bending σ_e, psi	$b_e = \sigma_e/\sigma_t$	Creep bending σ_c, psi	$b_c = \sigma_c/\sigma_t$
19	+45.8	+100.2				
16	−43.5	−101.5	650	0.929	547	0.782
14	+23.8	+72				
15	−25.6	−74	515	0.735	434	0.620
12	+8.4	+40				
13	−11.6	−41	355	0.507	298	0.425

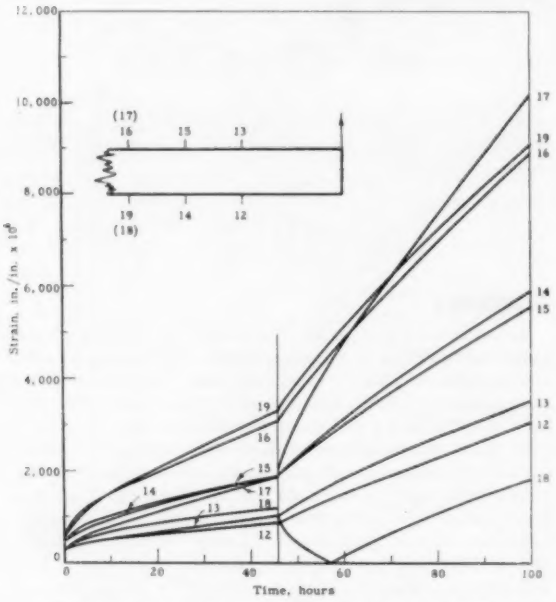


Fig. 2 Strain as a function of time for a tube in bending

Internal pressure added at 46 hours. Gages 17 and 18 lie in the tangential (circumferential) direction. The others lie in an axial direction. No distinction is made between tensile and compressive strain. Note that Gage 18 reverses sign at 56 hours.

($b_1^2 \ll 0.75$) and Equation (10) ($b_1^2 \gg 0.75$) are strictly applicable.

In Table 2 the ratio of axial creep rates with and without internal pressure is given. The average values from tension and compression gages may be compared with values predicted by Equations (9) and (10). This table shows very satisfactory agreement between experiment and analysis. For Location 12, 13, at which the bending stress is small relative to the pressure stress, we see that Equation (9) gives close agreement with the experimental value. As the bending stress increases, the ratio is overestimated by Equation (9) and agrees more closely with Equation (10).

Summary

The steady-state axial creep rate of a tube in bending may be greatly increased by the presence of internal pressure. When the maximum bending stress is small compared to the pressure stress, the bending stress under the combined loads can be shown very linearly over the cross section and thus may be calculated as if it were an elastic stress. As the bending stress increases, the bending stress distribution under combined bending and pressure approaches that corresponding to creep bending. For these two extreme cases, the factor by which the bending creep rate is increased by internal pressure is given by Equations (9) and (10), respectively. For any intermediate case such as Location 14, 15 of Tables 1 and 2, the factor lies between the values given by Equations (9) and (10).

Table 2 Ratio of axial creep rates with and without internal pressure

Location	$\frac{\text{Axial creep rate under bending and pressure}}{\text{Axial creep rate under bending only}}$ Experimental	Equation (9)	Equation (10)
19	2.19		
16	2.34		
Average 16, 19	2.26	2.41	2.31
14	3.02		
15	2.50		
Average 14, 15	2.76	2.87	2.56
12	4.76		
13	3.54		
Average 12, 13	4.15	4.13	3.30

References

1 A. E. Johnson, "Creep Under Complex Stress Systems at Elevated Temperatures," *Proceedings, I. Mech. E.*, vol. 164, 1951, p. 432.
2 E. L. Robinson, "Steam Piping Design to Minimize Creep Concentrations," TRANS. ASME, vol. 77, 1955, p. 1147.
3 E. P. Popov, "Stresses in Turbine Discs at High Temperatures," *Journal of The Franklin Institute*, vol. 243, 1947, p. 365.
4 R. W. Bailey, "The Utilization of Creep Test Data in Engineering Design," *Proceedings, I. Mech. E.*, vol. 131, 1935, p. 131.

DISCUSSION

K. R. Merckx[3]

The special case of creep bending of pressurized tubing under small bending moments may also be physically analyzed in the following manner: The axial strain and stress deviation of pressurized tubes are zero. The addition of small bending stresses, $\sigma_{sb} \ll \sigma$ where σ = flow criteria, will not appreciably alter the flow criteria, but will cause these axial strain or strain rates

$$\epsilon_s \text{ or } \dot{\epsilon}_s \simeq f(\sigma)\frac{2\sigma_{sb}}{3} = \frac{\sigma_{sb}}{E_{eff}},$$

where $f(\sigma)$ is any yielding function. The effective modulus E_{eff} can be experimentally determined as the ratio of the creep stress σ divided by the creep strain ϵ measured at the expected design life. Since for small bending stresses E_{eff} is a constant for given operating pressure, elastic methods of calculation can be used to determine relative bending deflections. This argument fails for bending stresses which are not small compared to the pressure stresses; thus methods similar to those described by the author must be used. The author's recognition of the effect of small bending stresses on creeping pressurized tubing is an important factor which is of use in many engineering applications.

[3] Hanford Laboratories Operation, General Electric Company, Richland, Wash. Mem. ASME.

J. T. TUCKER, JR.
Supervisor, Stress Analysis Group

E. E. COULTER
Section Head,
Engineering Physics Section
Assoc. Mem. ASME

L. F. KOOISTRA
Superintendent of Materials
Development. Mem. ASME

The Babcock & Wilcox Company
Research Center, Alliance, Ohio

Effect of Wall Thickness on Stress-Rupture Life of Tubular Specimens

Stress-rupture data obtained from tubular specimens stressed with internal pressure are compared with data from standard tension-bar specimens from the same heats of material. Agreement between the data of the two types of specimens is poor for thick-wall tubes when the hoop stress in the tube wall is calculated on an average stress basis. Better agreement is obtained with thin-wall tubes. These results show that any design formula that neglects the variation of stress through the thickness of a tube wall is unsatisfactory for thicker tubes.

A dimensionless parameter method is presented which provides satisfactory correlation of the data for three materials, four test temperatures, and a wide range of wall thickness. Several areas in which additional research would be of value are discussed.

Introduction

FOR THE DESIGN OF PRESSURE VESSELS and tubes that operate at high temperatures, stress-rupture is one of the more important properties of materials to be considered. Stress-rupture data indicate to the designer what service life can be expected of a material subjected to a specific stress and temperature before failure occurs. For this reason, stress-rupture properties of materials have been studied for many years, and abundant data on most commonly used materials are available in the literature.

Most of these data were obtained from cylindrical specimens machined from bar stock and loaded in pure tension, a condition which is rarely duplicated in actual service. These bar specimens do not account for metallurgical changes in the material which may occur during fabrication of parts, and the uniaxial tension stress does not duplicate the biaxial or triaxial stresses normally found in pressure vessels and tubes. Furthermore, although the initially applied stresses may be the same, the differences in the effect of creep on a bar specimen and on a tube can produce a difference in stress over a period of time.

For these reasons, the authors' company began a test program ten years ago to investigate the stress-rupture properties of materials in tubular form, stressed by internal pressure. Some of the early results have been reported in a previous paper.[1] The results of these continuing tests have proved to be quite fruitful and have provided information that could not be obtained by other methods. The tests have yielded information on the effect on stress-rupture life of physical and metallurgical discontinuities and stress concentrations, such as those occurring in welded tubing and tube bends. Also, since the tests included a wide range of outside-diameter-to-thickness ratios for several materials, the opportunity was afforded to observe the effect of tube wall thickness on stress-rupture life.

Obviously, the stress-rupture properties of the material are not changed as the wall thickness is varied. Therefore, the apparent change in rupture properties reflects the inability of the method of calculating stress to predict the stress condition that causes failure. It is shown that better correlation of bar specimens and tubular results can be obtained when the stress in the tube is calculated by a method that predicts a stress gradient in the tube wall, rather than assuming an average stress. The data for all wall thicknesses, materials, and temperatures covered in the tests are correlated reasonably well by use of dimensionless parameters. In this way all the results are compared and there is no need to resort to stress formulas which may not be theoretically applicable.

The combination of loads imposed on tubing in service varies considerably, depending on the design of the structure in which they are used. The data presented herein are applicable to only one common condition of loading, and further work is desirable to determine the effects of other combinations.

Materials and Apparatus

Three typical superheater materials were used in the tests reported in this paper:

1 Carbon steel, SA-210
2 Low alloy steel, SA-213, T22 (2¼Cr 1-Mo)
3 Austenitic steel, SA-213, TP321 (18-Cr 12-Ni Ti)

Their physical properties, chemical compositions, heat-treatments, and ASTM specifications are listed in Tables 1, 2, and 3. It is apparent that the specimens were representative of their respective specifications. All of the tubing of each type and the bar specimens with which they are compared were taken from the same heat of material and were heat-treated together so that their physical properties and microstructure were similar.

Tubular Specimens. A typical tubular stress-rupture specimen is shown in Fig. 1. It consists of a 17-inch length of tubing whose

[1] L. F. Kooistra, R. U. Blaser, and J. T. Tucker, Jr., "High-Temperature Stress-Rupture Testing of Tubular Specimens," TRANS. ASME, vol. 74, 1952, pp. 783–791.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 1, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 24, 1959. Paper No. 59—Met-11.

Nomenclature

S = average hoop stress, psi
S_m = maximum hoop stress, psi
S_R = stress required for rupture in a bar specimen in a specified time, psi
P = internal pressure, psig
P_T = internal pressure required for rupture in a tubular specimen in a specified time, psig
D_i = inside diameter, inches
D_o = outside diameter, inches
t = tube wall thickness, inches

Table 1 Physical properties

Material	Heat no.	Ultimate strength, psi	Yield point, psi	Elongation, per cent
2¼-Cr 1-Mo thick-wall tube	H1913L	68700	37000	35
2¼-Cr 1-Mo thin-wall tube	924	65100	47300	54
2¼-Cr 1-Mo ASTM Spec.	...	60000	25000	30
Type 321 thick-wall tube	20526	80900	50000	71
Type 321 ASTM Spec.	...	75000	30000	35
SA-210 thick-wall tube	T1088	63000	43500	65
SA-210 ASTM Spec.	...	60000	37000	25

Table 2 Chemical composition, (per cent)

Material	Heat No.	C	Mn	P	S	Si	Cr	Ni	Mo	Ti
2¼-Cr 1-Mo thick-wall tube	H1913L	0.09	0.41	0.022	0.018	0.29	2.12	...	0.98	...
2¼-Cr 1-Mo thin-wall tube	924	0.10	0.39	0.020	0.019	0.27	2.18	...	0.97	...
2¼-Cr 1-Mo ASTM Spec.	...	0.15 max	0.30–0.60	0.030 max	0.030 max	0.50 max	1.90–2.60	...	0.87–1.13	...
Type 321 thick-wall tube	20526	0.065	1.79	0.010	0.025	0.42	17.92	12.36	...	0.38
Type 321 ASTM Spec.	...	0.080 max	2.00 max	0.030 max	0.030 max	0.75 max	17.00–20.00	9.00–13.00	...	1.00 max
SA-210 thick-wall tube	T1088	0.232	0.46	0.010	0.013	0.33	Small residuals		...	...
SA-210 ASTM Spec.	...	0.27 max	0.80 max	0.058 max	0.048 max	0.10 min	...	...	...	...

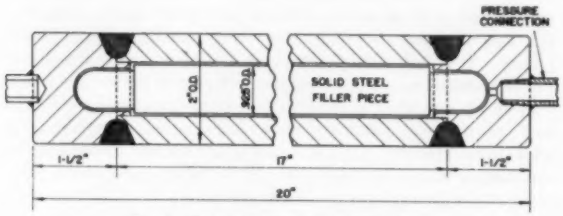


Fig. 1 Tubular stress-rupture specimen

Table 3 Heat-treatment

Material	Heat no.	Heat-treatment
2¼-Cr 1-Mo	H1913L	1550 F for 1 hour 25 F per hour furnace cooled to 1200 F
2¼-Cr 1-Mo	924	Furnace cooled (maximum rate) to room temperature
Type 321	20526	1900 F for ½ hour Water quench
SA-210	T1088	1650 F for ½ hour Air cooled to below 400 F 1300 F for ½ hour Air cooled to room temperature

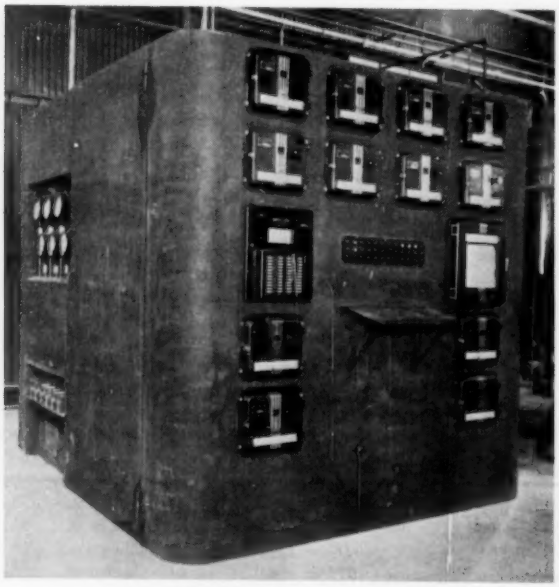


Fig. 2 Exterior view of stress-rupture apparatus

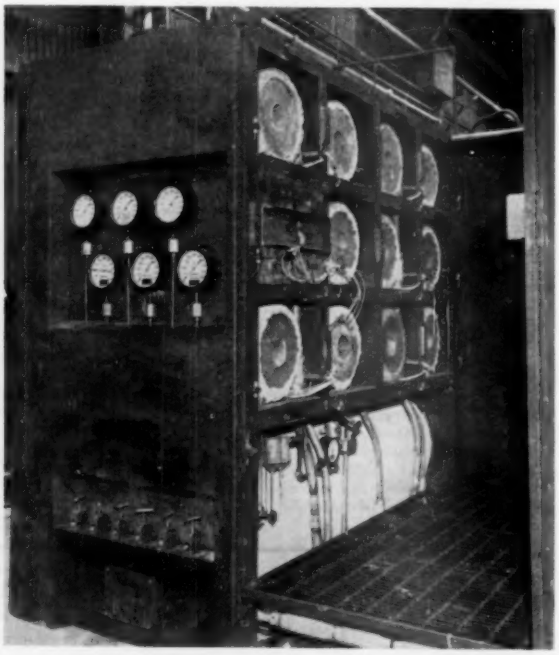


Fig. 3 Interior view of stress-rupture apparatus

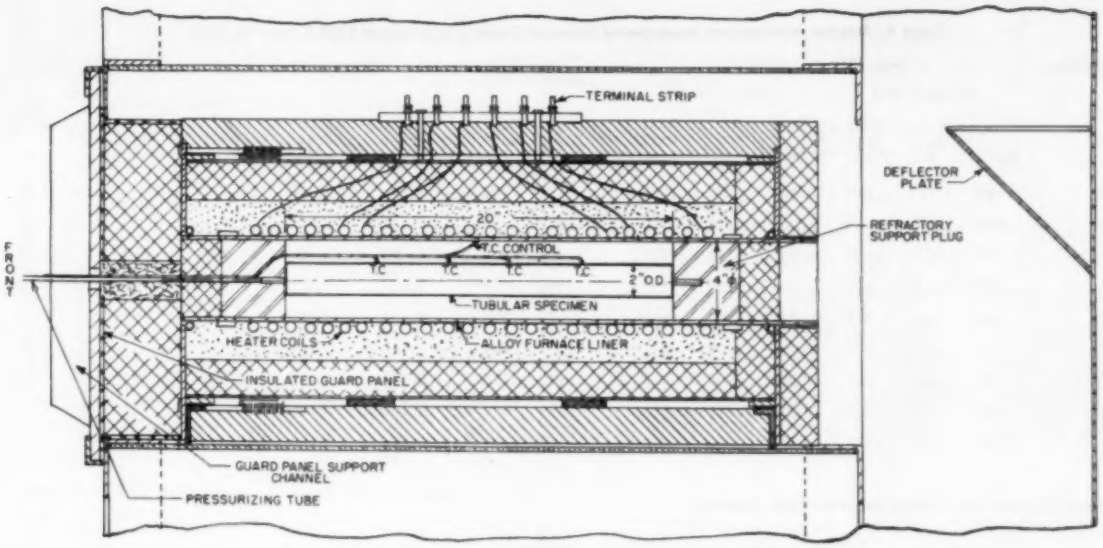


Fig. 5 Arrangement of one unit of multiple-station stress-rupture apparatus

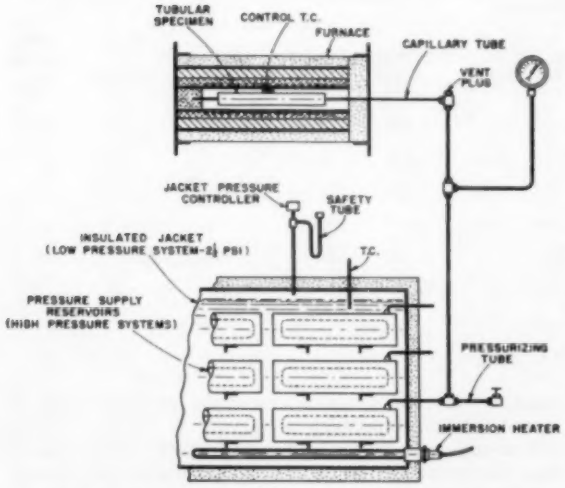


Fig. 4 Typical section through one unit of stress-rupture apparatus

ends are closed by welding on closure heads. Before the heads are welded, a solid steel filler piece is inserted into the tube to reduce the internal volume and thus minimize the explosive forces when rupture occurs. For most tests, the specimens were made from standard commercial tubing with no machined surfaces or other refinements; however, the specimens used to study the effect of varying the diameter-to-thickness ratios were accurately machined on the inside and outside surfaces to eliminate eccentricity and produce a uniform wall thickness. Short lengths of $^{9}/_{16}$-in. OD tubing are welded to the closure heads to support the specimen in the center of the cylindrical furnace. A $^{1}/_{8}$-in. OD tube attached to one of the support tubes connects the specimen to the pressurizing system. Five Chromel-Alumel thermocouples are located along the specimen to provide a basis for controlling the temperature of the specimen and the temperature distribution along its length.

Apparatus. The tubular stress-rupture apparatus is illustrated in Figs. 2 and 3. It contains twelve separate stations where tubular specimens are stressed with internal steam pressure in a controlled temperature furnace until a stress-rupture failure results. Located in the front section of the apparatus are the instrument panel and the individual power panels. These units control the saturable reactors that supply continuous power to the constant temperature furnaces.

The furnaces are positioned horizontally in their insulated cubicles at the rear of the apparatus. A sectional view, Fig. 4, shows a tubular specimen, with its instrumentation, installed in its furnace ready for testing. The electric furnace, with its separately controlled zones, will heat the specimen evenly over the central ten inches of the specimen, maintaining a maximum temperature gradient over this length of ±4 F and holding temperature variations at any point within the gage length to ±3 F. An insulated guard panel closes the front of the cubicle and causes the force and fragments of rupture to be directed toward the back of the apparatus where heavy deflector plates eliminate any hazard to operating personnel.

The rear section of the apparatus also contains the high pressure circuits used to pressurize the specimens to the desired stress level. As shown on Fig. 5, each circuit contains a high-pressure reservoir which is held at a constant temperature (220 F) by its electrically heated low-pressure water jacket. The reservoir, which is the major volume of each circuit, utilizes its elastic dilation and the compressibility of the water to minimize the effect of specimen creep and small temperature fluctuations on the specimen loading pressure. Internal specimen pressures, as indicated by the gages at the sides of the apparatus, are maintained within ±25 psig by periodic additions of make-up water from an air-driven water pump.

All of the tension bar specimens were tested with standard stress-rupture equipment in which the specimens are stressed at elevated temperature by a dead weight lever system until rupture occurs.

Test Procedure. Since there is some eccentricity in most commercial seamless tubing, it becomes necessary to establish some method of describing the dimensions of the specimens. Therefore, before the end closures are welded on a specimen, its initial dimensions are measured with a micrometer. Outside diameters are measured on 90-degree planes at $2^{1}/_{2}$-in. intervals to detect any large irregularities along the specimen. If large irregularities

Table 4 Tubular stress-rupture properties of thin-wall Croloy 2¼ tubing at 1100 F, heat No. 924

Specimen No.	Actual Tube Dimensions						Calculated Dimensions		Testing Conditions			Results	
	Stamped End			Opposite End									
	O.D. In.	Min. Wall In.	Max. Wall In.	O.D. In.	Min. Wall In.	Max. Wall In.	Avg. I.D. In.	Min. Wall In.	Inter. Press Psi	Hoop* Stress Psi	Lamé Stress Psi	Time to Rupture Hr.	Cir. Elong. %
1	2.004	.206	.219	2.003	.211	.215	1.577	.209	5,570	21,100	24,200	15	26
2	1.998	.212	.222	2.000	.213	.229	1.563	.212	4,680	17,200	19,800	51	28
3	2.004	.216	.229	1.997	.193	.235	1.565	.205	3,340	12,800	14,600	372	18
4	1.998	.214	.227	1.994	.218	.227	1.556	.216	3,300	11,900	13,700	580	15
5	1.999	.210	.217	1.995	.200	.220	1.576	.205	3,120	12,000	13,700	858	15
6	1.996	.205	.223	1.998	.207	.227	1.567	.206	2,600	9,900	11,300	4,348	15
7	2.006	.209	.225	2.002	.208	.220	1.571	.208	2,770	10,400	12,000	4,786	14
8	2.003	.193	.223	2.000	.210	.223	1.579	.202	2,800	11,000	12,500	6,422	13
9	2.006	.204	.219	1.999	.201	.222	1.579	.203	2,280	8,900	10,200	7,928	17

*Hoop Stress Calculated By Thin-Tube Formula

Table 5 Tubular stress-rupture properties of 2¼-Cr 1-Mo tubing at 1100 F, heat No. H19131

Specimen No.	Actual Tube Dimensions						Calculated Dimensions		Testing Conditions			Results	
	Stamped End			Opposite End									
	O.D. In.	Min. Wall In.	Max. Wall In.	O.D. In.	Min. Wall In.	Max. Wall In.	Avg. I.D. In.	Min. Wall In.	Inter. Press Psi	Hoop* Stress Psi	Lamé Stress Psi	Time to Rupture Hr.	Cir. Elong. %
1	2.497	.560	.607	2.497	.560	.596	1.336	.56	7,470	8,900	13,700	394	30
2	2.496	.572	.594	2.496	.575	.580	1.336	.574	7,670	8,900	13,900	507	40
3	2.496	.577	.584	2.495	.571	.588	1.336	.574	6,030	7,000	10,900	1,744	33
4	2.496	.562	.609	2.495	.561	.602	1.330	.561	5,600	6,600	10,300	4,098	38
5	2.496	.573	.590	2.494	.571	.590	1.333	.572	5,130	6,000	9,300	5,314	26
6	2.497	.431	.437	2.496	.432	.436	1.628	.432	5,300	10,000	13,200	299	29

*Hoop Stress Calculated By Thin-Tube Formula

are found, the specimen is rejected. Tube wall thickness is measured at each end of the specimen at eight points 45 degrees apart. From this information, the point of minimum thickness at each end is located. The outside diameter at this point and the opposite wall thickness are then measured and used to calculate the inside tube diameter. The dimensions of the two ends are averaged to give an average inside diameter and an average minimum wall thickness for each specimen. These two average dimensions describe a concentric hypothetical tube and are used to calculate the internal pressure required to cause rupture within the desired test period.

After a specimen is assembled, it is evacuated and filled with distilled water. The vacuum makes it possible to fill the specimen completely with water through the small-diameter pressurizing tubing. The specimen and necessary insulation and guards are then installed in a furnace of the test apparatus.

Following installation, the specimen is hydrostatically tested and then its temperature is raised to the desired value. During the temperature-raising period, a ten-foot standpipe is connected to the pressurizing system. This arrangement maintains a low pressure on the system, which prevents the high-pressure reservoir from boiling dry, and allows the water in the specimen to expand. When the specimen temperature has stabilized, the pressure is increased slowly by an air-driven pump until the test pressure is attained. The temperature and pressure are maintained at the proper levels throughout the life of the specimen by making slight adjustments as required.

When a specimen fails, it is allowed to cool and then is removed from the apparatus. Its dimensions are recorded and it is examined metallurgically.

Comparison of Stress-Rupture Data. The physical properties, chemical analyses, and heat-treatments of the three materials used in these test series are listed in Tables 1 through 3, and have already been discussed. Tables 4 through 8 list the specimen dimensions, test conditions, and results of all tests. It is evident from these summary tables that only certain facts are accurately known about a tubular stress-rupture specimen; namely, its material, its dimensions, the test temperature, the test pressure, and the time for rupture. While these facts are necessary to make good use of the results, one must still answer the important question of how shall the stress in the tube wall be computed.

All tubes that are subjected to internal pressure at low temperatures are known to have a higher hoop stress at the inside surface than at the outside, provided the stresses are kept within the elastic limit. The difference between the stresses on the inside and outside surfaces is equal to the internal pressure. Plastic theory also predicts a stress gradient across the tube wall, but with the highest stress being on the outer surface. At elevated temperatures, it has often been assumed that, after a short period of time, these stress differences become smaller because of higher

Table 6 Tubular stress-rupture properties of thick-wall Type 321 tubing at 1200 F, heat No. 20526

Specimen No.	Actual Tube Dimensions						Calculated Dimensions		Testing Conditions			Results	
	Stamped End			Opposite End									
	O.D. In.	Min. Wall In.	Max. Wall In.	O.D. In.	Min. Wall In.	Max. Wall In.	Avg. I.D. In.	Min. Wall In.	Inter. Press Psi	Hoop* Stress Psi	Lamé Stress Psi	Time to Rupture Hr.	Cir. Elong. %
1	2.003	.514	.521	2.003	.513	.522	.968	.514	15,000	14,100	24,200	527	1.3
2	2.003	.509	.525	2.003	.506	.528	.969	.508	12,780	12,200	20,800	1,393	4.3
3	2.003	.515	.521	2.003	.515	.518	.969	.515	11,120	10,500	18,200	2,279	---
4	2.003	.514	.520	2.003	.514	.521	.969	.514	10,490	9,900	17,000	3,199	6.0
5	2.003	.514	.522	2.003	.514	.523	.967	.514	9,870	9,300	16,000	3,683	6.0
6	2.003	.509	.526	2.003	.513	.521	.969	.511	9,190	8,700	14,900	5,389	8.0
7	2.003	.510	.526	2.003	.512	.525	.967	.511	8,580	8,100	13,900	6,675	7.0
8	2.003	.507	.528	2.003	.507	.527	.969	.507	7,910	7,600	12,900	8,972	9.4

*Hoop Stress Calculated By Thin Tube Formula

Table 7 Tubular stress-rupture properties of Type SA-210 tubing, heat No. T1088 and Type 321 tubing, heat No. 20526

Spec. No.	Actual Tube Dimensions						Calculated Dimensions		Testing Conditions				Results	
	Stamped End			Opposite End										
	O.D. In.	Min. Wall In.	Max. Wall In.	O.D. In.	Min. Wall In.	Max Wall In.	Avg. I.D. In.	Min. Wall In.	Temp. °F	Inter. Press Psi	Hoop* Stress Psi	Lame Stress Psi	Time to Rupture Hr.	Cir. Elong. %
Type A-210														
1A	1.940	.466	.468	1.940	.466	.467	1.007	.466	850	17,590	19,000	30,600	220	54
2B	1.941	.375	.375	1.941	.375	.375	1.191	.375	850	11,960	19,000	26,400	768	37
3C	1.941	.249	.251	1.941	.250	.250	1.441	.249	850	6,570	19,000	22,800	2,221	38
4D	1.941	.156	.157	1.941	.156	.156	1.629	.156	850	3,630	19,000	21,000	3,551	43
5A	1.940	.462	.465	1.940	.463	.463	1.014	.462	950	10,030	11,000	17,600	622	65
6B	1.940	.377	.377	1.941	.377	.377	1.187	.377	950	6,980	11,000	15,300	1,250	68
7C	1.941	.249	.252	1.941	.251	.251	1.440	.25	950	3,820	11,000	13,200	2,734	50
8D	1.941	.156	.156	1.941	.156	.157	1.628	.156	950	2,110	11,000	12,100	4,472	52
Type 321														
9A	1.939	.467	.468	1.940	.467	.468	1.004	.467	1200	13,950	15,000	24,200	646	7
10B	1.939	.372	.374	1.939	.373	.373	1.193	.373	1200	9,360	15,000	20,800	879	12
11C	1.940	.246	.248	1.940	.247	.247	1.446	.246	1200	5,110	15,000	17,900	871	6
12D	1.935	.154	.156	1.935	.155	.156	1.625	.154	1200	2,850	15,000	16,500	945	5
13a	1.940	.468	.470	1.940	.468	.469	1.003	.468	1200	14,000	15,000	24,200	1,663	9
14b	1.941	.374	.376	1.941	.375	.376	1.191	.374	1200	9,430	15,000	20,800	972	–
15c	1.942	.251	.251	1.942	.251	.252	1.439	.251	1200	5,230	15,000	18,000	2,115	9
16d	1.942	.156	.157	1.941	.157	.157	1.628	.157	1200	2,880	15,000	16,600	1,452	7

*Hoop Stress Calculated By Thin-Tube Formula

Table 8 Stress-rupture data from tension-bar specimens

Specimen no.	Diameter, in.	Stress, psi	Hours to failure	Elongation, per cent
Material: 2¼-Cr 1-Mo, heat No. 924, 1100 F				
1	0.252	17000	107	21
2	0.252	17000	140	48
3	0.254	17000	191	41
4	0.254	15000	281	20
5	0.252	15000	458	64
6	0.253	13000	1175	50
7	0.254	13000	1486	28
8	0.253	12000	2363	38
9	0.253	12000	3255	25
10	0.253	10000	7179	14
11	0.253	10500	7788	38
Material: 2¼-Cr 1-Mo, heat No. H1913L, 1100 F				
1	0.503	15000	121	62.5
2	0.503	13500	498	80.0
3	0.504	11500	4267	85.5

Specimen no.	Diameter, in.	Stress, psi	Hours to failure	Elongation, per cent
Material: 18-Cr 12-Ni Ti (Type 321), heat No. 20526, 1200 F				
1	0.322	20000	572	21.0
2	0.322	17000	1517	26.0
3	0.320	16000	2098	14.0
4	0.315	14000	3317	25.5
5	0.321	12000	7999	18.0
Material: SA-210, heat No. T1088, 850 F and 950 F				
850 F 1	0.321	28000	238	62.0
850 F 2	0.322	23000	1988	47.0
850 F 3	0.321	21500	4772	78.0
850 F 4	0.321	20000	12801	78.5
950 F 1	0.321	17000	330	54.5
950 F 2	0.321	15000	1388	85.5
950 F 3	0.321	13500	2998	82.0
950 F 4	0.320	12000	6228	69.0

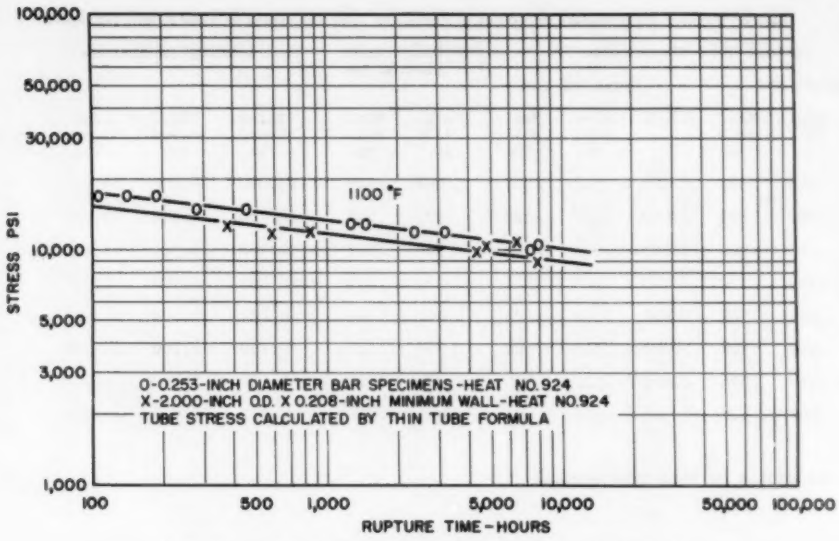


Fig. 6 Rupture data for 2¼-Cr 1-Mo tubing; thin tube versus tension bar

creep rates in the more highly stressed locations. Such an equalization of stress across a tube wall would permit estimating tube life on the basis of an average hoop stress as calculated by the thin tube formula

$$S = P \frac{D_i}{2t}$$

where

S = average hoop stress, psi
P = internal pressure, psig
D_i = inside diameter, inches
t = tube wall thickness, inches

Hoop stresses calculated by this formula are used in this paper as a basis for comparing tubular stress-rupture data with bar data.

The first series of tests was made on specimens from two-inch tubing with 0.203-in. minimum wall thickness of 2¼-Cr 1-Mo material. The results of these tests at 1100 F are shown in Fig. 6 along with the results of bar specimens from the same heat of material. Although the tubular specimens failed at a stress slightly lower than that of the bar specimens for any given rupture time, the agreement is reasonably good. This difference in stress is typical for relatively thin-wall tubing and is probably due to a basic difference in the effect of creep on the actual stress in the material of the two types of specimen. The nominal stress value of both specimens is calculated from the load and the dimensions of the specimen at the beginning of a test. As creep occurs, the load on a bar specimen remains constant and the cross-sectional area decreases somewhat, thereby increasing the true stress in the material. However, a double effect of creep on a tubular specimen takes place, a reduction of wall thickness or load bearing area and an increase in inside diameter which represents the area on which the internal pressure acts. This produces a vicious circle in which each increment increases the other throughout the life of a specimen.

The next test series was made with specimens from 2½-in. tubing with 0.570-in. minimum wall thickness of 2¼-Cr 1-Mo material. The results of these tests at 1100 F are shown in Fig. 7. It is evident that the tubular data fall well below the bar data from the same heat of material when the stresses in the tubes are computed using the thin tube formula. Of special significance is the fact that one tubular specimen, which was bored and reamed to a 0.427-in. wall thickness, failed at a stress higher than would be indicated by the data of the thicker-wall tubes. It should also be noted from the physical properties, chemical analyses, and stress-rupture properties of the bar specimens, that this heat of 2¼-Cr 1-Mo material is very similar to the previous heat, making possible a direct comparison of three specimens with different wall thicknesses whose stresses are calculated by the same formula. This comparison points to a relationship between wall thickness and rupture stress whereby the thickest tube at a given stress level has a much shorter rupture life than the thinner tubes.

It seems obvious that the stress-rupture properties of the material itself are not affected to any great extent by variations in wall thickness. It follows, therefore, that the assumption of average stress in the tube wall is invalid (at least for the temperatures studied), and that the stress causing failure in the thick-wall tubes must be appreciably greater than the average value.

Better agreement between the tubular and bar data is obtained when the maximum hoop stress in the tubular specimens is calculated using the Lamé formula

$$S_m = P \frac{D_o^2 + D_i^2}{D_o^2 - D_i^2}$$

where

S_m = maximum hoop stress, psi
P = internal pressure, psig
D_o = outside diameter, inches
D_i = inside diameter, inches

This comparison is also shown in Fig. 7 for the thick-wall specimens. The success of this formula in providing better correlation between tubular and bar stress-rupture data is thought to be purely empirical, since it was derived by applying the elastic theory, and stress-rupture is essentially a plastic phenomenon. Furthermore, the Lamé formula predicts the maximum stress at the inner surface, while an inspection of the specimens after failure indicates that the failure, in nearly every case, originated

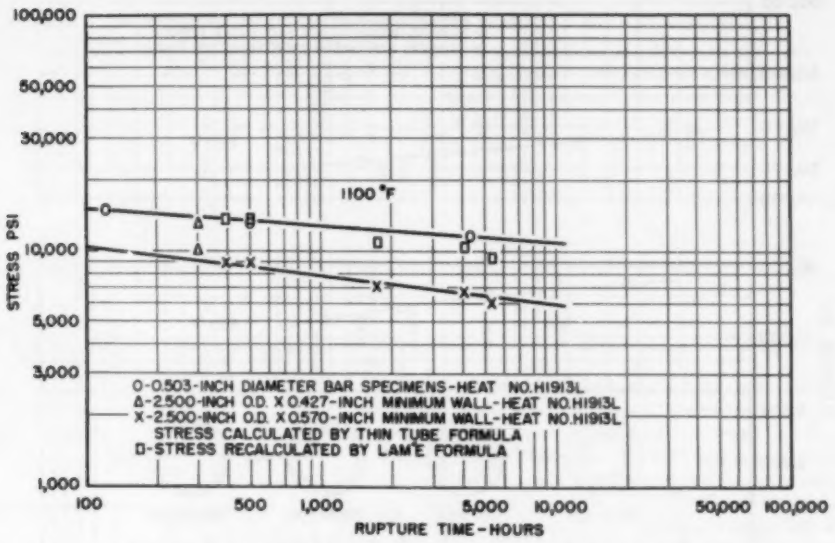


Fig. 7 Rupture data for $2^{1}/_{4}$-Cr 1-Mo tubing; thick-wall tube versus tension bar

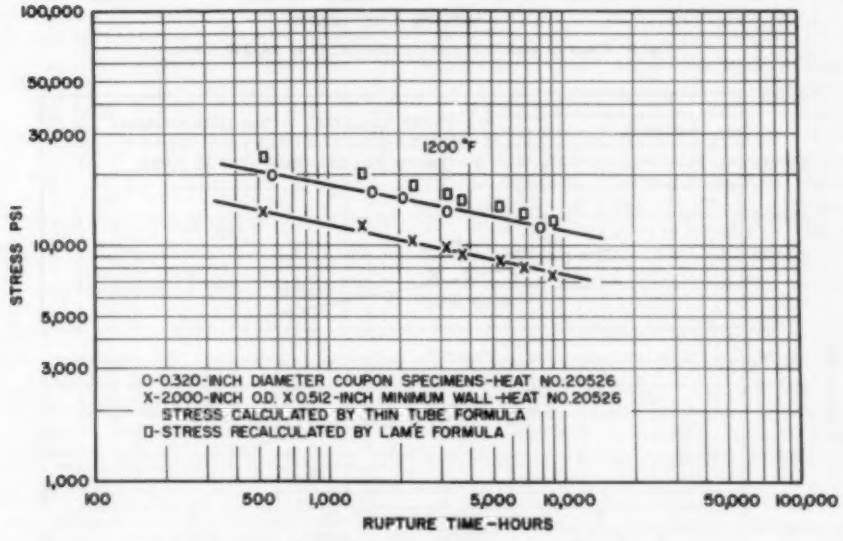


Fig. 8 Rupture data for Type 321 tubing; thick-wall tube versus tension bar

on the outside surface. Therefore, the Lamé formula is not presented necessarily as a recommended method of calculating stresses in tubes at high temperatures, but simply to illustrate the desirability of using a method of calculating stress which includes the effect of a stress gradient across the tube wall and produces a higher than average stress value. However, it will be shown that regardless of any theoretical reasoning, the Lamé formula will give a reasonable correlation between tubular data and bar data over a wide range of materials, temperatures, and wall thicknesses.

The third set of data were obtained at 1200 F from two-inch tubes with 0.512-in. minimum wall thickness of Type 321 stainless steel (18-Cr 12-Ni Ti). As shown in Fig. 8, the agreement between tubular and bar data is again poor when the stresses in the tubular specimens are computed by the thin-tube formula. The data, although somewhat on the high side, are in better agreement when the tube stresses are calculated according to the Lamé formula.

These results pointed out the need for more carefully controlled tests that would cover a greater range of diameter-to-thickness ratios. The specimens for these tests had accurately machined inside and outside surfaces. They were made of two-inch diameter tubes of SA-210 steel, and the wall thicknesses were varied from 0.16 to 0.48 in. Test temperatures of 850 and 950 F were selected. The test pressures were chosen so that the stresses in all specimens were the same, as calculated by the thin-tube formula. Therefore, if this method of calculating stress were valid, all specimens would rupture in approximately the same length of time. As shown in Fig. 9, the specimens did not fail in the same time period, the rupture times varying by more than 200 to one. At each temperature, the specimen with the thickest wall failed in the shortest time, and the rupture time increased

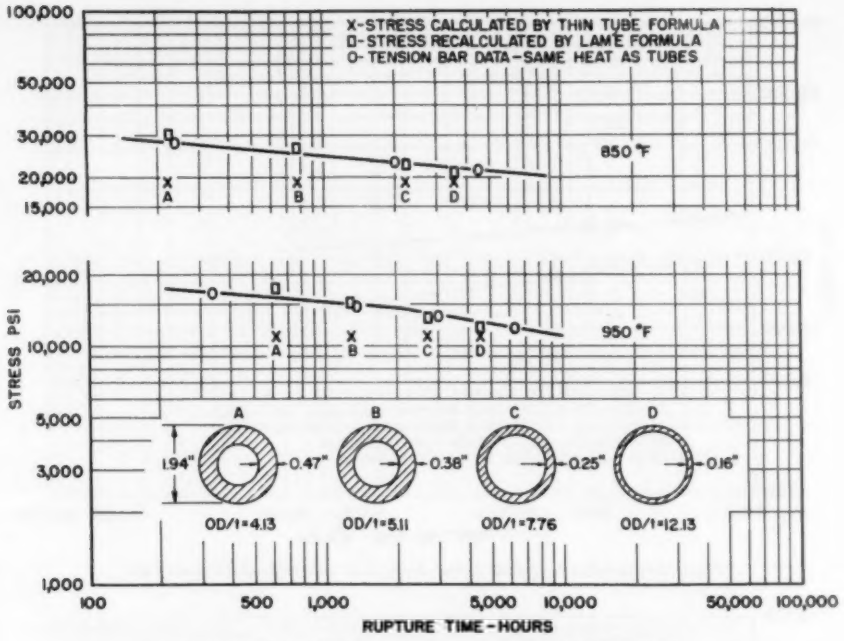


Fig. 9 Effect of wall thickness on rupture life of SA-210 tubing

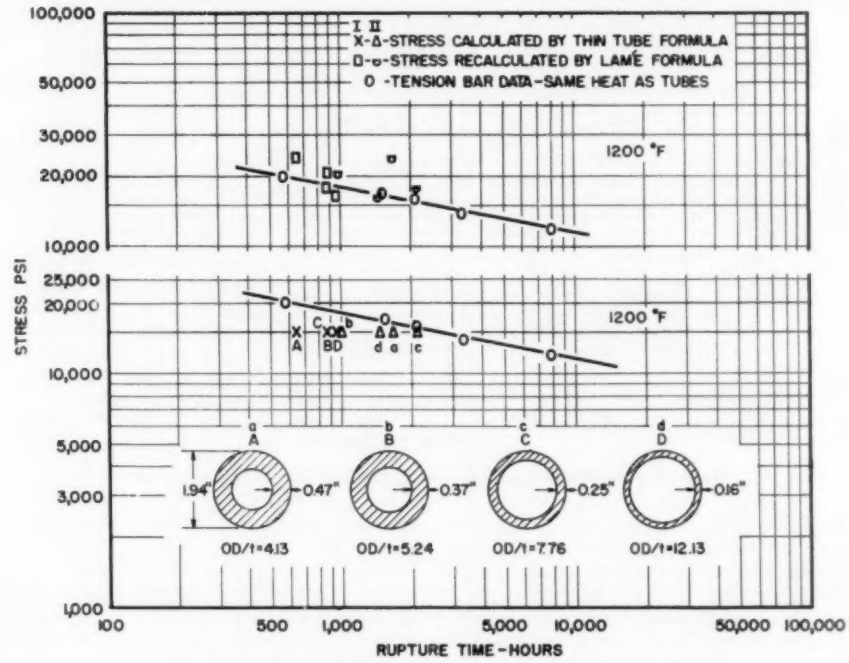


Fig. 10 Effect of wall thickness on rupture life of Type 321 alloy tubing

as the wall thickness decreased. When the stresses in the tubular specimens are calculated using the Lamé formula, the data fall on a straight line on log-log paper with the slight negative slope characteristic of stress-rupture data. These results agree closely with those of tension-bar specimens of the same material.

The results of two series of similar tests on Type 321 stainless steel material at 1200 F are shown in Fig. 10. The first set of specimens, which were stressed at 15,000 psi by the thin-tube formula, all failed short of the bar data with the thickest wall specimen rupturing in the shortest time and the thinnest in the longest time. It should be noted that the total spread of the data is much less than for the steel tubing and that the thinner tubes are bunched close together. However, the same general tendencies, as mentioned previously, are present but are not so

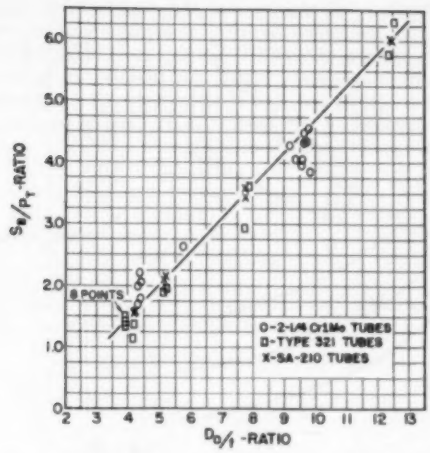


Fig. 11 Correlation of tubular data by dimensionless parameters

well defined. The second set of these specimens also fell short of the bar data and in a random order that is difficult to explain. A metallurgical investigation of the tubing from which these specimens were made revealed a marked difference in appearance and nonhomogeneity of grain size across the wall thickness. The machining of various amounts of metal from the inside of the tubes to provide the various wall thicknesses could have resulted in specimens with somewhat different properties. This may account for some of the data scatter experienced. These tests will be repeated on Type 304 tubing which has a more uniform grain structure and will provide more uniform specimens.

Although a very large number of both theoretical and empirical formulas have been proposed for calculating the behavior of tubes under high temperature and pressure, none seems to account completely for results which have been obtained experimentally. Furthermore, since stress-rupture properties of materials can be obtained with less expense from bar specimens than from tubular specimens, and since most of the existing data were obtained with bar specimens, it is desirable to establish a relationship between these data and tube performance so that the data can be better utilized for design purposes. In an attempt to establish such a relationship without resorting to the use of formulas, the correlation shown in Fig. 11 was devised. This correlation uses two dimensionless parameters, D_o/t and S_B/P_T, to relate tubular and tension bar data. The parameter D_o/t, which is the ratio of the outside diameter of the tube to wall thickness, describes the geometry of the tube. The parameter S_B/P_T is, in effect, a stress concentration or loading factor. The term S_B is the stress required to produce failure of a tension-bar specimen in a given time. P_T is the pressure necessary to cause failure of a tubular specimen of the same material and at the same temperature in the same given time. It is assumed that the stress-rupture properties of a material are the same in both tubular and tension-bar specimens and that the stress in the tube which causes failure is directly proportional to internal pressure and a function of the geometry parameter D_o/t.

The degree of correlation of the data in Fig. 11 indicates that these basic assumptions are correct, since the data for all materials and all test temperatures fall along a single line. However, a considerable amount of scatter is noted in some cases. Some of this scatter may be attributed to test inaccuracies and material variations which would normally be expected. It appears, however, that some of this apparent scatter may be due to an effect of time. For example, the data from the thick-wall tubes of Type 321 alloy, all of which have approximately the same D_o/t ratio, show that the S_B/P_T ratio values tend to increase as the time to rupture increases as indicated:

S_B/P_T	Time to rupture, hr
1.37	527
1.33	1393
1.39	2279
1.38	3199
1.48	3683
1.43	5389
1.47	6675
1.50	8972

The result of this tendency may be that a precise comparison of data from the two types of specimens is not possible. Further investigation of this question is needed.

It is well to realize that the test condition imposed on these tubes at constant temperature and pressure does not represent all possible conditions of service. The axial stress was caused by internal pressure alone. In many designs tubes act not only as pressure containers but also as structural parts, and the axial stress may approach or even exceed the hoop stress. The effect of larger axial loads on the stress-rupture properties of tubes is another area that requires further investigation.

Still another such area is the effect of heat transfer on stress-rupture life of tubes. Most tubes are used as heat transfer surfaces and, therefore, have a temperature differential between inside and outside surfaces. As a result, thermal stresses must be added to the stresses caused by pressure, and the properties of the material will vary through the tube wall because of the temperature variation. Whether or not these phenomena will have any significant effect on tube life can only be determined experimentally, and we hope to investigate these questions in the near future.

Summary

The foregoing stress-rupture tests, although not including the effects of heat transfer and structural loading, permit the following observations to be made:

1 The variation in rupture time for tubes of the same material over a range of wall thicknesses indicates that the assumption of average stress across the tube wall is invalid for thick-wall tubes at least for the temperature tested.

2 The average stress assumption is satisfactory for relatively thin-wall tubes where the effect of stress gradient across the tube wall is negligible.

3 The Lamé formula for maximum stress at the inside surface seems to correlate the various materials over a range of wall thickness and temperature. This formula should be considered empirical as the assumptions under which it was derived are no longer valid.

4 The dimensionless parameter S_B/P_T and D_o/t correlate all the results to a linear relationship with some evidence that the scatter shown is time dependent.

Acknowledgment

The authors wish to acknowledge the co-operation of W. E. Leyda and the Creep Laboratory at The Babcock & Wilcox Company Research Center in producing the bar data used in this paper.

DISCUSSION

W. L. Harding[2]

The authors are to be commended for this presentation of a considerable amount of stress-rupture data on tubular specimens. It is particularly valuable to have the detailed test measurements reported in such a form that a variety of possible correlations can be investigated.

The writer's company agrees with the authors' belief in the value of testing specimens of tubular form, and has such a program in progress. Tubular stress-rupture equipment is in many respects similar to that described in this paper, but differs in that inert gas rather than steam is used as the pressurizing medium. Results from this testing are not yet complete enough to be included in this discussion. What data are available show a trend in general agreement with that of the authors, and indicate that presently used correlations of tubular rupture with uniaxial rupture tests may become less satisfactory as the tube thickness-to-diameter ratio increases. The authors have done valuable service in calling attention to the uncertainties in this area. The situation emphasizes the need for much more test data on thick-walled tubes. In view of the scatter of the present data it is felt that effort to develop a revised formula or correlation would be premature. When more data are available it is suggested that the various analyses described by Burrows, Michel, and Rankin[3] may be a useful starting point for consideration of a modified formula, if the additional data show that a change is needed.

The authors' Fig. 11 plots the ratios (S/P) versus (D_o/t), and the straight line they draw from these data is given approximately by the expression

$$(S/P) = 0.545\,(D_o/t) - 0.73$$

A trial of the widely used expression for pipe wall thickness,

$$S/P = 0.5\,(D_o/t) - 0.4$$

gives a line which appears to fit the plotted points about as well. The disturbing aspect of Fig. 11 from the designer's standpoint is the wide spread in the (S/P) ratio for the thick-walled tubes. For the relatively thin-walled tubes, of D_o/t of 9 or 10, the range from highest to lowest (S/P) is only about 20 per cent, which is a tolerable scatter. For thick-walled tubes of D_o/t of 4 to 4.5, however, the highest (S/P) is about twice the lowest.

Cause of this amount of scatter needs to be resolved by further testing before any simple relation between (S/P) and (D_o/t) can be considered to be proved adequate for thick-walled tubes.

J. F. Traexler[4] and P. C. Warner[4]

The authors' idea of correlating creep-rupture data obtained from tubular specimens with data obtained from uniaxial creep-rupture tests is of great practical interest since it indicates one way of reducing creep-rupture testing expense and complexity while at the same time insuring adequate tube or cylinder design. A successful correlation and understanding would, to a large degree, eliminate the need for the more complicated tubular testing in so far as tube or cylinder design is concerned

Since there is as yet no generally accepted fracture theory, the question as to what stress, or combination of stresses, should be considered when correlating fracture data is an important one. It is somewhat unfortunate that the authors have chosen to consider only the maximum principal stress to be of importance, and have thereby eliminated the important triaxiality effect. This might have been included, for example, by considering both the maximum principal stress and the spherical or hydrostatic component of the stress tensor. The general trend in such a case would be as shown in Fig. 12.

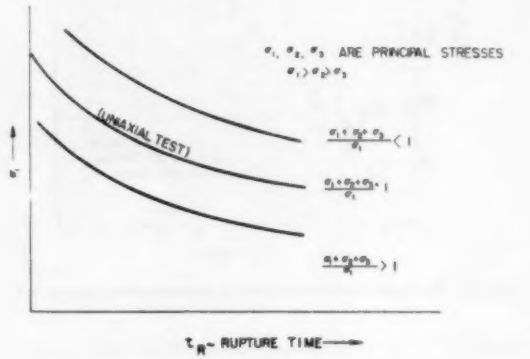


Fig. 12

The authors' tests may be grouped roughly into two classes; one in which the geometry remained fixed and the stress level varied, and the second in which the geometry was varied and the membrane stress level remained fixed. In both types of tests, the authors' claim that, if the maximum principal stress in the tubular test is calculated from the known internal pressure and geometry using an elastic analysis, better correlation is obtained with uniaxial data than if a membrane analysis is made. It is again unfortunate that the authors have limited themselves to either a membrane or an elastic analysis and make no mention of a plastic analysis, which is at least theoretically correct for the loadings and temperatures involved.

The writers have made such an analysis for the authors' tests on type 321 tubing assuming a nonwork-hardening material. Fortunately, the actual stress distribution is not too sensitive to this assumption and small error is introduced. A further argument for the plastic analysis is that it not only apparently predicts the correct magnitude of the peak principal stress but also predicts the proper type of distribution. Whereas the peak elastic tangential stress occurs at the bore, the peak plastic tangential stress occurs at the outside, and the authors have pointed out that all failures originated at the outside. The results are shown in Figs. 13 and 14 which correspond respectively to Figs. 8 and 10 in the paper. Fig. 13 shows the results of tests on tubes of the same geometry with various internal pressures. In all cases the curves signify calculated stresses. The uniaxial data are shown for purposes of comparison.

In assessing Fig. 13, it is obvious that the plastic analysis gives much better agreement with the uniaxial data than either the elastic or the membrane analysis. The question as to whether the plastic analysis gives a perfect fit within the limits of experimental error cannot be answered however, since in the present case there is no objective way to assess the experimental error. This suggests that taking cognizance of some of the more elementary aspects of experimental design from a statistical standpoint would greatly enhance the utility of the otherwise very careful experimental work done by the authors. For example, if, instead of testing at eight pressures, the authors had tested two specimens at each of four pressures, independent estimates of the experimental variability and lack of fit of the data to any model could be made. By comparing one variability with the

[2] Research and Product Development, Combustion Engineering, Inc., New York, N. Y. Mem. ASME.

[3] W. R. Burrows, R. Michel, and A. W. Rankin, "A Wall-Thickness Formula for High-Pressure High-Temperature Piping," TRANS. ASME, vol. 76, 1954, pp. 427–444.

[4] Development Engineering, Steam Division, Westinghouse Electric Corporation, Lester, Pa. Assoc. Mem. ASME.

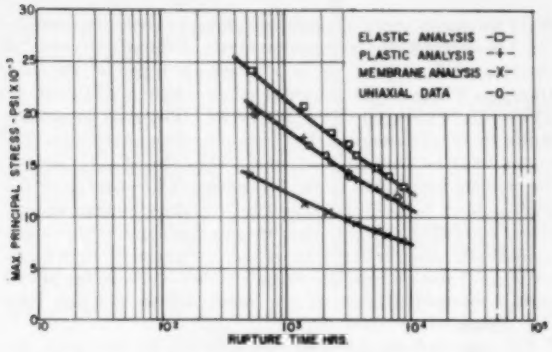


Fig. 13 Type 321 tubing, 1200 F

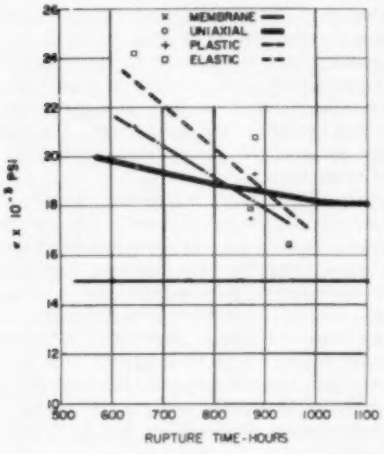


Fig. 14

other, it would be possible to conclude whether or not any particular model was consistent with the data. Fig. 14 shows the results of tests on tubes of various geometries but with the same membrane stress level. Here, the correlation is decidedly poorer than for the previous tests, but again the plastic analysis agrees more closely with the uniaxial data than does the elastic analysis. The authors have pointed out that machining may have caused changes in material properties and that specimens used in these latter tests showed marked nonhomogeneity in so far as grain size was concerned. This may be a part of the reason for the poor correlation in these cases.

The writers have also studied the authors' Fig. 11 with some interest and have plotted the theoretical curves obtained from membrane, elastic, and plastic analyses in Fig. 15. Assuming, as the authors did, that perfect correlation exists (i.e., $\sigma_T = S_B$) the plastic analysis gives the following expression:

$$\frac{S_B}{P_i} = \frac{1}{\ln\left(\frac{D_o/t}{D_o/t - 2}\right)}$$

It can be seen that this curve gives a good fit to the data and at the same time has theoretical significance.

A perusal of the curves in Fig. 15 indicates that, if part of the reason for the tests is to choose among elastic, membrane, and plastic-stress analyses, the testing should be concentrated toward the thick-walled tube end of the diagram in order that the theories give results sufficiently different to enable a choice to be made

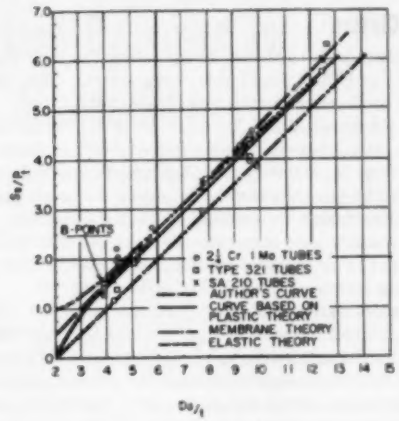


Fig. 15

experimentally. Further, if the purpose of testing included a reasonable extrapolation of results or the answering of some other like question, experimental-design techniques would permit an optimum design of the experiments for the chosen purposes and the economics of the case. No less important, they permit an objective assessment of the chance of drawing erroneous conclusions from the data.

A. B. Wilder[5]

The authors have presented important high-temperature data with specific reference to the behavior of small diameter heavy wall tubing. Due to higher steam temperatures and pressures there has been an increased use of heavy wall tubing in boilers. It is our understanding that most of this heavy wall tubing is used under Section I of the ASME Boiler and Pressure Vessel Code. The thin-tube formula used by the authors is similar to the Boiler Code formula. The Lamé formula provides for heavier wall material and the results of the authors indicate this formula should be used. Data of the type presented by the authors and additional information concerning various heats of the same grade and different grades of steel should be available from various laboratories before definite conclusions are formed. We would appreciate comments of the authors regarding the use of heavier wall tubing in the ASME Boiler Code.

Has consideration been given to determination of the creep rate under conditions of the tubular creep-rupture test? Although tubes fail by rupture, their behavior before rupture in comparison with bar tests would be of interest.

The authors have indicated that in the creep-rupture tests of tubing the inside diameter increases and the wall thickness decreases as the tube expands during testing. In burst testing large diameter pipe 24 to 36 in. OD and 40 ft long at ambient temperatures, it has been observed that for 75,000-psi ultimate strength material the bursting strength as calculated by Barlow's formula is approximately 5,000 psi less than the transverse tensile strength. This has been associated in part with the increase in diameter of the pipe as we are in the plastic rather than elastic range during the latter stages of burst testing. The wall thickness decreases during the test and this results in work-hardening which would be in the direction of increasing the strength properties of the material. Comments of the authors regarding the possibilities of work-hardening or change in the microstructure of the material during the reduction of wall thickness at elevated temperatures would be of interest.

[5] Chief Metallurgist, National Tube Division, United States Steel Corporation, Pittsburgh, Pa.

Authors' Closure

The authors wish to thank the discussers for their thoughtful comments and their valuable suggestions. The questions raised will be discussed in the order presented.

The participation of Mr. W. L. Harding's company in testing tubular specimens is very welcome indeed, because the additional data will help to definitely establish trends and reduce uncertainties due to scatter inherent in testing materials properties. Their results, though incomplete, support the authors' position that the presently used correlation between bar data and tubular data may not be satisfactory for thick-wall tubes. This problem is now being reviewed by an ASME code committee.

The authors agree with Mr. Harding's comment that data scatter in the plot of ratios of S_B/P_i versus D_o/t, particularly in the thick wall range, leaves much to be desired. However, the curve was drawn principally through the crossed points for the SA-210 specimens whose dimensions were accurately machined inside and out. A recent set of data, obtained on accurately machined specimens of Type 304 alloy, also fits the plotted curve closely. These confirming results point to the possibility that some of the scatter shown may be due to the use of unmachined specimens. Nevertheless, the authors agree that this scatter needs to be reduced or explained before this simple relationship can be used for thick-wall tube design.

The authors wish to express their appreciation to Messrs. Traexler and Warner for their discussion of the plastic analysis we referred to in our paper. They have stated the problem exactly when they question what stress or combination of stresses should be used when correlating fracture data, especially since there is no generally accepted fracture theory. Since this statement describes the present state of the art, any procedure which correlates the test data without taking into account strain hardening, creep, relaxation, wall thinning, differential diametral growth, increased pressure area, combined stress conditions that vary across the tube wall, varying creep rate with varying combined stress conditions, initial elastic and plastic deformation, oxidation, metallurgical changes, etc., must be considered empirical. A satisfactory theoretical analysis of the many variables affecting stress-rupture life in a tube may never be attained. However, a number of empirical relationships may be set up which will be adequate for tube design. The good agreement shown by the discussers in their plastic analysis of the Type 321 tubing must be considered an empirical relationship because of the reasons mentioned in the foregoing. This same analysis, when applied to the thick-wall $2^1/_4$-Cr 1-Mo specimens gives a poorer correlation. These qualifications also apply to the elastic analysis presented by the authors and it remains for future correlations of test data to determine whether either of the procedures mentioned, or perhaps some other, will be best suited for tube design.

The suggested use of a statistical approach for setting up the test programs is a useful one as it permits an optimum design of the experiment. However, when only eight specimens of a given material are to be tested, the advantage of determining a probable experimental error by testing two specimens each at four test conditions may be overshadowed by testing each specimen at a different stress level and drawing a least squares curve through the eight points. The slope of a stress-rupture curve is better defined by eight points than by four. If a larger number of specimens are involved, the statistical procedure suggested can be used to an advantage.

With respect to Mr. Wilder's question regarding the use of heavier wall tubing in the ASME Boiler Code, our results indicate a smaller factor of safety for thick-wall tubes than for thin wall tubes when using the code formula. A code committee has been set up to review the data pertaining to this problem.

Creep rates have not been determined under the conditions of the tubular stress-rupture tests. However, two tests were made by periodically measuring at room temperature the diametral growth occurring in a specimen during a cyclic stress-rupture test.

The authors feel that work-hardening and changes in microstructure occur during stress-rupture testing with a resulting effect on rupture life.

Design of Pressurized Cylinders for High-Temperature Applications

J. F. TRAEXLER
Development Engineering,
Westinghouse Electric Corporation,
Lester, Pa. Assoc. Mem. ASME

General equations for the stresses in a thick-walled cylinder in a state of plane strain are derived considering "steady-state" creep. A specific form of the creep-rate function is assumed and numerical examples are included to show the effect of geometry and material properties.

Introduction

THE EVER-INCREASING USE of higher pressures and temperatures in the power-generation field has led to a renewal of interest in the mechanical design of components subjected to creep. One commonly occurring problem concerns the design of thick-walled pressurized cylinders carrying radial temperature gradients where the stresses and temperatures are such as to cause creep. Clearly an elastic analysis is meaningless in such cases. This paper will present a stress analysis for such a case assuming the cylinder to be fully plastic and the pressure and thermal loading to be independent of time.

The first problem confronting the design engineer deals with the choice of a suitable failure criterion. In some cases, components fail because of excessive distortion, whereas for other applications fracture itself constitutes failure. The choice of a failure criterion obviously depends on the application of the component and the consequences of failure. There is however one feature common to both of these failure criteria. Each requires a detailed knowledge of the stresses before any conclusions as to the adequacy of a design can be reached. Consequently, this paper will deal in the main with the problem of determining stresses as functions of the space co-ordinates, with only a brief discussion of the procedure for applying either of the failure criteria.

Simply stated, the method will consist of extending the fundamental equations of plasticity to cover the creep problem, recognizing that creep is both temperature and stress dependent.[1] The derivations will be general at first and specific engineering formulas will be obtained from the more general solutions. Numerical examples will be included to show the effect of material properties, temperature gradients, and geometry on the stress distribution.

[1] C. D. Weir, "The Creep of Thick Tubes Under Internal Pressure," *Journal of Applied Mechanics*, vol. 24, TRANS. ASME, vol. 79, 1957, pp. 464–466.

Contributed by the Metals Engineering Division and presented at the Metals Engineering Conference, Albany, N. Y., April 29–May 3, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 24, 1959. Paper No. 59—Met-8.

Discussion

The problem to be considered deals with a heavy-walled cylinder in a state of generalized plane strain (ϵ_z = const). It will be assumed that elastic strains are negligible in comparison to plastic strains, thus allowing the use of the Saint Venant equations of plastic flow in place of the more rigorously correct equations of Reuss. Furthermore, it will be assumed that only long-time problems are of interest, thus allowing the use of a constant creep rate. In other words, second-stage creep will be assumed to predominate. Triaxial stress states will be reduced to equivalent uniaxial states by means of the distortion strain energy theory of von Mises-Hencky.

A brief outline of the method of solution follows: For greater detail the reader is referred to the Appendix.

The Saint Venant laws are:

$$\dot{\epsilon}_{ij}' = \Psi \sigma_{ij}' \tag{1}$$

$$\epsilon_{ij}'' = 0 \tag{2}$$

where

$$\Psi = \Psi(\sigma_e, T) \tag{3}$$

and in accordance with the distortion strain energy theory

$$2\sigma_e^2 = (\sigma_t - \sigma_r)^2 + (\sigma_t - \sigma_z)^2 + (\sigma_r - \sigma_z)^2 \tag{4}$$

Equation (1) states the existence of a functional relationship between the strain-rate deviator and the stress deviator while (2) is the incompressibility relation. Equation (3) specifies that the function relating strain rate and stress is dependent on both the state of stress (σ_e) and the state of temperature (T).

Making use of these equations and the previously stated assumptions, we arrive at the following general expressions for the radial and tangential stresses

$$\sigma_r = -\frac{1}{\sqrt{3}} \left\{ \int \left(1 + \frac{\sigma_e}{\Psi} \frac{\partial \Psi}{\partial \sigma_e} \right) d\sigma_e + \int \left(\frac{\sigma_e}{\Psi} \frac{\partial \Psi}{\partial T} \frac{dT}{dr} \right) dr \right\} + C \tag{5}$$

Nomenclature

ϵ = strain	Ψ = creep-rate function	$(\ )_t$ = tangential
$\dot{\epsilon}$ = strain rate, hr^{-1}	θ = creep-rate function at constant T	$(\ )_z$ = axial
σ = stress, psi	λ = creep-rate function at constant σ	$(\ )_e$ = combined or equivalent
T = temperature, deg F		$(\ )_a$ = at bore
r = radius, in.	A, B, n, α = material constants	$(\ )_b$ = at outside
a = inside radius, in.	**Subscripts**	**Superscripts**
b = outside radius, in.	$(\ )_r$ = radial	$(\ )'$ = deviator part of tensor
P = internal pressure, psig		$(\ )''$ = spherical part of tensor
u = radial displacement, in.		

$$\sigma_t = \frac{-1}{\sqrt{3}} \left\{ \left(1 + \frac{\sigma_e}{\Psi} \frac{\partial \Psi}{\partial \sigma_e}\right) r \frac{d\sigma_e}{dr} + \frac{r\sigma_e}{\Psi} \frac{\partial \Psi}{\partial T} \frac{dT}{dr} + \int \left(1 + \frac{\sigma_e}{\Psi} \frac{\partial \Psi}{\partial \sigma_e}\right) d\sigma_e + \int \frac{\sigma_e}{\Psi} \frac{\partial \Psi}{\partial T} \frac{dT}{dr} dr \right\} + C \quad (6)$$

If one now assumes

$$\Psi(\sigma_e, T) = \theta(\sigma_e) \cdot \lambda(T) \quad (7)$$

there is considerable reduction and (5) and (6) become

$$\sigma_r = \frac{-1}{\sqrt{3}} \{(1 + \ln \theta + \ln \lambda)\sigma_e - \int \ln \lambda \, d\sigma_e - \int \ln \theta \, d\sigma_e\} + C \quad (8)$$

$$\sigma_t = \frac{-1}{\sqrt{3}} \left\{(1 + \ln \theta + \ln \lambda)\sigma_e + \frac{r\sigma_e}{\theta} \frac{d\theta}{dr} + \frac{r\sigma_e}{\lambda} \frac{d\lambda}{dr} + \frac{d\sigma_e}{dr} - \int (\ln \theta + \ln \lambda) \, d\sigma_e \right\} + C \quad (9)$$

Finally, if one assigns specific forms to $\theta(\sigma_e)$ and $\lambda(T)$, as for example

$$\theta(\sigma_e) = A\sigma_e^n \quad (10)$$

$$\lambda(T) = Be^{\alpha T} \quad (11)$$

and recalls that for a cylinder with a steady-state radial temperature gradient

$$T = m \ln r + k \quad (12)$$

where

$$m = \frac{T_a - T_b}{\ln \left(\frac{a}{b}\right)} \quad (13)$$

and

$$k = \frac{T_b \ln a - T_a \ln b}{\ln \frac{a}{b}} \quad (14)$$

the resulting equations are

$$\sigma_r = -P \left\{ \frac{\left(\frac{a}{r}\right)^{\frac{2+\alpha m}{1+n}} - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}{1 - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}} \right\} \quad (15)$$

$$\sigma_t = P \left\{ \frac{\frac{1 - n + \alpha m}{1 + n} \left(\frac{a}{r}\right)^{\frac{2+\alpha m}{1+n}} + \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}{1 - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}} \right\} \quad (16)$$

Note that for the isothermal case $m = 0$ and if, in addition, $n = 0$, we arrive at the isothermal elastic case.

The form of $\theta(\sigma_e)$ used in Equation (10) is but one of several common in engineering usage. The form of $\lambda(T)$ chosen in (11) is not quite so common. Originally introduced by Bailey[2] it has lost favor in recent years due to the application of chemical rate theory to creep processes. The form of (11) consistent with chemical rate theory would be

$$\lambda(T) = De^{\frac{-Q}{RT}} \quad (17)$$

Either this form or that in (11) can be made to fit test data reasonably well provided that the temperature range is not too large. The greatest advantage of the form used in (11) is that it provides considerable mathematical simplification and leads to final closed form equations. For the range of temperature of practical interest, little error is introduced by using it in place of (17), and there is a great saving of computational time and effort.

Numerical examples based on Equations (15) and (16) are shown in Figs. 1 to 4. In each case the thin cylinder or membrane approximation is included for purposes of comparison. The values of n and α used are not for any specific material but are typical values that might be expected in practice.

Fig. 1 shows the effect of temperature gradient on the stress distribution for a cylinder with fixed geometry and material. In all cases the bore temperature is held constant. Note that there is a shift of the stress field with the lower temperature and, hence, stronger elements at the outside picking up load at the expense of the inner fibers. It is this feature that makes double wall construction desirable for high-stress, high-temperature applications. Note also the closeness of the membrane theory, especially in the absence of a radial thermal gradient. No appreciable difference in the radial stress distribution occurs for the gradients chosen.

Fig. 2 shows the effect of n on the stress distribution in a given cylinder with a fixed gradient. Higher values of n designate lower strain-hardening ability and this results in flatter tangential stress distributions. This is to be expected since an n of zero corresponds to the elastic case. For high n values, the membrane theory improves as an approximation. Higher n values, however, mean that there will be less of a shift toward the stronger, cooler elements and for an optimum design a balance must be worked out between loading, geometry, and material. Again, no appreciable difference exists in the radial stress distribution for the range of n values chosen.

Fig. 3 shows the effect of α on the stress distribution. Equations (15) and (16) show that an increase in α has the same effect on the stress distribution as a decrease in n and vice versa. This trend is borne out by the curves.

Fig. 4 shows the effect of a geometry change for a given material and temperature gradient. As may be expected from membrane theory considerations, thinner cylinders show higher over-all stress levels.

Conclusions

General equations for the radial and tangential stresses in a thick-walled pressurized cylinder in a state of generalized plane strain have been derived. Engineering formulas have been arrived at by specifying a specific form of the creep-rate function. The effects of material properties, geometry, and thermal loading have been brought out through the use of numerical examples.

Having the complete stress distribution, it is fairly obvious as to the procedure to be followed for designs using rupture as a failure criterion. Experimentally determined creep to rupture data are required for the stress and temperature levels involved. From this, an estimate of the service life may be made.

For designs having specified limits of distortion, an additional integration of Saint Venant's strain-rate equations with respect to time is necessary. The geometric definition of strain is then applied to compute the actual displacements.

[2] R. W. Bailey, "The Utilization of Creep Test Data in Engineering Design," The Institution of Mechanical Engineers, London, England, 1935, p. 246.

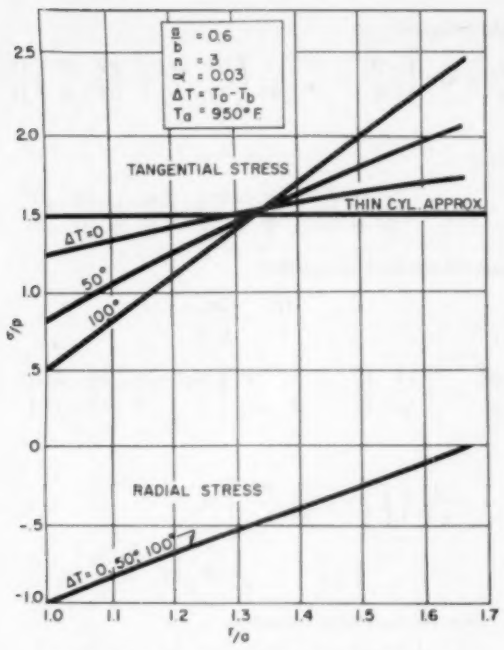


Fig. 1 Effect of thermal gradient on stress distribution

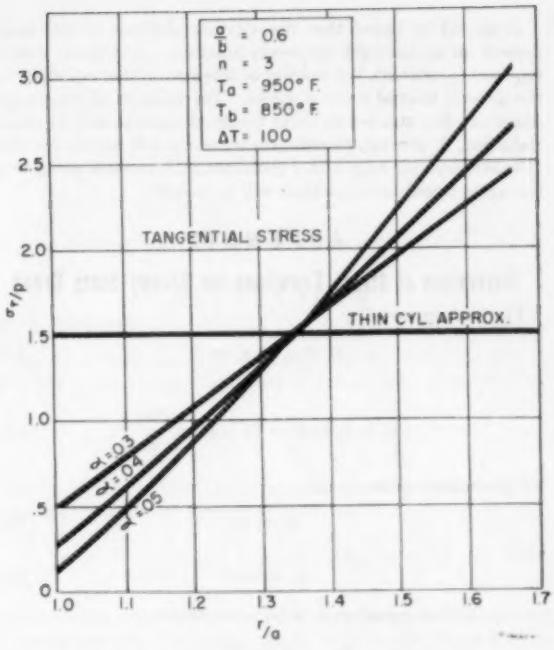


Fig. 3 Effect of material constant (α) on stress distribution

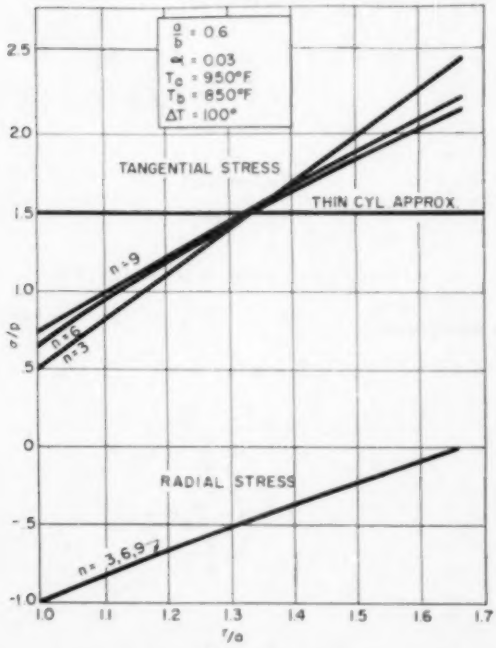


Fig. 2 Effect of material constant (n) on stress distribution

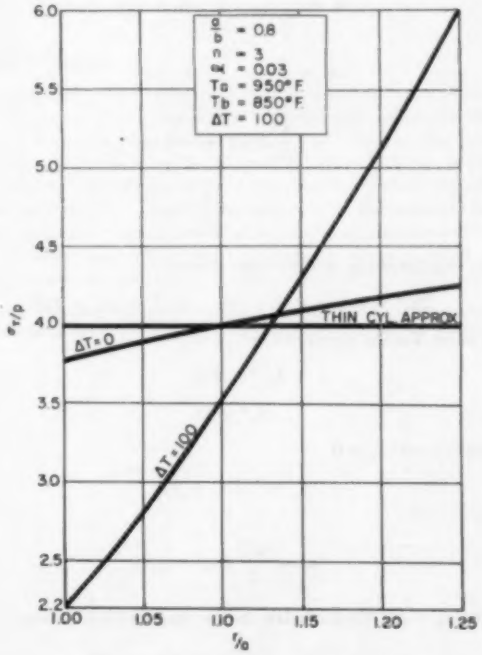


Fig. 4 Effect of geometry on stress distribution

It should be noted that the solution methods of this paper depend on an assumed creep-rate function. Additional work is required to establish the validity of this assumption especially for the general triaxial state of stress. The solution of the present paper can be extended to cover transient cases as well as steady state but, in general, closed-form solutions will not be possible. It is also possible to consider problems with variable creep rates but again closed-form solutions will not result.

APPENDIX

Derivation of Stress Equations for Steady-State Creep

For incompressibility

$$\epsilon_r + \epsilon_t + \epsilon_z = 0 \tag{18}$$

or

$$\dot{\epsilon}_r + \dot{\epsilon}_t + \dot{\epsilon}_z = 0; \quad (\dot{\epsilon}) = \frac{d(\epsilon)}{dt} \tag{19}$$

for generalized plane strain

$$\dot{\epsilon}_z = 0 \tag{20}$$

then

$$\dot{\epsilon}_r = -\dot{\epsilon}_t \tag{21}$$

the equilibrium equation in polar co-ordinates is

$$\frac{d\sigma_r}{dr} + \frac{\sigma_r - \sigma_t}{r} = 0 \tag{22}$$

the geometric definition of strain gives

$$\epsilon_r = \frac{du}{dr} \qquad \dot{\epsilon}_r = \frac{d\dot{u}}{dr} \tag{23}$$

$$\epsilon_t = \frac{u}{r} \qquad \dot{\epsilon}_t = \frac{\dot{u}}{r} \tag{24}$$

$$\frac{d}{dr}(r\dot{\epsilon}_t) = -\dot{\epsilon}_t \tag{25}$$

$$\dot{\epsilon}_t = \frac{c}{r^2} \tag{26}$$

from the distortion strain-energy theory

$$2\sigma_e^2 = (\sigma_t - \sigma_r)^2 + (\sigma_r - \sigma_z)^2 + (\sigma_z - \sigma_t)^2 \tag{27}$$

the Saint Venant equations are:

$$\dot{\epsilon}_{ij}{}' = \Psi\sigma_{ij}{}' \tag{28}$$

$$\epsilon_{ij}{}'' = 0 \tag{29}$$

using this with $\dot{\epsilon}_z = 0$

$$\sigma_z = {}^1/_2(\sigma_r + \sigma_t) \tag{30}$$

then

$$\sigma_e = \frac{\sqrt{3}}{2}(\sigma_t - \sigma_r) \tag{31}$$

Applying Saint Venant's law to the tangential directions

$$\dot{\epsilon}_t = \Psi\sigma_t{}' = \Psi\frac{1}{2}(\sigma_t - \sigma_r) = \frac{c}{r^2} \tag{32}$$

then

$$\sigma_e = \frac{\sqrt{3}C}{r^2\Psi} \tag{33}$$

differentiating

$$\frac{d\sigma_e}{dr} = \sqrt{3}\left\{\frac{-2C}{r^3\Psi} - \frac{C}{r^2}\left[\frac{1}{\Psi^2}\left(\frac{\partial\Psi}{\partial\sigma_e}\frac{d\sigma_e}{dr} + \frac{\partial\Psi}{\partial T}\frac{dT}{dr}\right)\right]\right\} \tag{34}$$

finally

$$\sigma_e = -\frac{r}{2}\frac{d\sigma_e}{dr}\left(1 + \frac{\sigma_e}{\Psi}\frac{\partial\Psi}{\partial\sigma_e}\right) - \frac{\sigma_e}{\Psi}\frac{r}{2}\frac{\partial\Psi}{\partial T}\frac{dT}{dr} \tag{35}$$

using the equilibrium equation

$$\frac{d\sigma_r}{dr} = \frac{2\sigma_e}{\sqrt{3}r} \tag{36}$$

$$\frac{d\sigma_r}{dr} = -\frac{1}{\sqrt{3}}\left\{\left[1 + \frac{\sigma_e}{\Psi}\frac{\partial\Psi}{\partial\sigma_e}\right]\frac{d\sigma_e}{dr} + \frac{\sigma_e}{\Psi}\frac{\partial\Psi}{\partial T}\frac{dT}{dr}\right\} \tag{37}$$

or

$$\sigma_r = -\frac{1}{\sqrt{3}}\left\{\int\left[1 + \frac{\sigma_e}{\Psi}\frac{\partial\Psi}{\partial\sigma_e}\right]d\sigma_e + \int\frac{\sigma_e}{\Psi}\frac{\partial\Psi}{\partial T}\frac{dT}{dr}\,dr\right\} + C \tag{38}$$

and from the equilibrium equation

$$\sigma_t = r\frac{d\sigma_r}{dr} + \sigma_r \tag{39}$$

now if

$$\Psi = \theta(\sigma_e)\,\lambda(T) \tag{40}$$

$$\sigma_r = -\frac{1}{\sqrt{3}}\{(1 + \ln\theta + \ln\lambda)\sigma_e - \int\ln\lambda\,d\sigma_e - \int\ln\theta\,d\sigma_e\} + C \tag{41}$$

if

$$\begin{aligned}\theta(\sigma_e) &= A\sigma_e^n \\ \lambda(T) &= Be^{\alpha T}\end{aligned} \tag{42}$$

$$\sigma_r = -\frac{1}{\sqrt{3}}\{(1 + n)\sigma_e + \int\alpha\sigma_e dT\} + C \tag{43}$$

for steady state

$$T = m\ln r + k \tag{44}$$

$$m = \frac{T_a - T_b}{\ln\frac{a}{b}} \tag{45}$$

$$k = \frac{T_b\ln a - T_a\ln b}{\ln\frac{a}{b}} \tag{46}$$

but

$$\sigma_e = \frac{\sqrt{3}C}{r^2\Psi} = \left[\frac{\sqrt{3}C}{AB}\right]r^{-2}\sigma_e^{-n}e^{-\alpha T} \tag{47}$$

$$r = e^{\left(\frac{T-k}{m}\right)} \tag{48}$$

finally

$$\sigma_e = \left[\frac{\sqrt{3}C}{AB}\right]^{\frac{1}{1+n}} e^{\frac{2k}{m(1+n)}} e^{-\left(\frac{2+\alpha m}{m(1+n)}\right)T} \tag{49}$$

or

$$\sigma_e = De^{-qT} \tag{50}$$

then

$$\sigma_r = -\frac{1}{\sqrt{3}}\left\{(1+n)De^{-qT} + \int \alpha De^{-qT}dT\right\} + C \tag{51}$$

$$\sigma_r = -\frac{De^{-qT}}{\sqrt{3}}\left(1 + n - \frac{\alpha}{q}\right) + C \tag{52}$$

but D and C are determined from

$$\sigma_r = -P \text{ @ } r = a \tag{53}$$
$$\sigma_r = 0 \text{ @ } r = b$$

finally

$$\sigma_r = -P\left\{\frac{\left(\frac{a}{r}\right)^{\frac{2+\alpha m}{1+n}} - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}{1 - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}\right\} \tag{54}$$

and from the equilibrium equation

$$\sigma_t = P\left\{\frac{\frac{1-n+\alpha m}{1+n}\left(\frac{a}{r}\right)^{\frac{2+\alpha m}{1+n}} + \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}{1 - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}\right\} \tag{55}$$

$$\sigma_e = \frac{\sqrt{3}}{2}P\left\{\frac{\frac{2+\alpha m}{1+n}\left(\frac{a}{r}\right)^{\frac{2+\alpha m}{1+n}}}{1 - \left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}}}\right\} \tag{56}$$

note also that since $0 < \frac{a}{b} < 1$

$$\left(\frac{a}{b}\right)^{\frac{2+\alpha m}{1+n}} = 1 \quad \text{only if} \quad \frac{2+\alpha m}{1+n} = 0$$

Applying L'Hospitals' rule

$$\lim_{\left(\frac{2+\alpha m}{1+n}\right)\to 0} \sigma_r = -P\left(\frac{\ln b/r}{\ln b/a}\right) \tag{57}$$

Thus showing that the stresses always remain finite.

DISCUSSION

K. R. Merckx[3]

An exponential dependence upon flow criteria may also be handled in the manner suggested by the author. Assuming

$$\dot{\epsilon}_{ij}' = A \exp\left[\frac{\sigma_e}{\sigma_0} + \alpha(T-k)\right]\sigma_{ij}'$$

with the temperature distribution given by Equation (12), the resulting equation for the equivalent stress is

$$\sigma_e = \frac{\sqrt{3}\,P}{2\ln(b/a)} + \left(\frac{2+\alpha m}{2}\right)\sigma_0 \ln\left[(b/a)/(r/a)^2\right].$$

In this model, very low strain rates are represented by negative exponentials, which is not physically realistic (the use of a negative exponential is equivalent to having tensile strains being caused by compressive stresses). Thus this material model cannot be used unless

$$P/\sigma_0 > \frac{(2+\alpha m)}{\sqrt{3}}\left[\ln\left(\frac{b}{a}\right)\right]^2.$$

Author's Closure

The author wishes to express his appreciation to Dr. Merckx for his interest in the paper and his comments on an alternate form of $\theta(\sigma_e)$. In view of the form of $\lambda(T)$ used in the paper, the form of $\theta(\sigma_e)$ used by Dr. Merckx is particularly convenient. It is also very gratifying to see some attention paid to the often neglected question of physical significance of the assumed form of flow criteria as distinguished from mathematical convenience. This is a point that cannot be emphasized too strongly and should always be investigated before applying the results of any analysis.

[3] Fuel Element Design Operation, Hanford Laboratories Operation, General Electric Company, Richland, Wash. Mem. ASME.

R. P. BENEDICT
Development Engineer, Steam Division, Westinghouse Electric Corporation, Lester, Pa.
Assoc. Mem. ASME

The Response of a Pressure-Sensing System

A first-order, first-degree, linear expression is developed to describe the transient behavior of a practical pressure-sensing system. The development is in terms of usual fluid-flow parameters.

An experiment is described whereby the transient behavior is determined for a variety of pressure-sensing systems. Experimental uncertainties are determined.

Comparisons are drawn between several published expressions, the proposed linear expression, and the experimental data reported herein, to indicate the validity of the proposed expression.

Selected references to related work are given.

Pressures are usually sensed by means of probes which consist of surface orifices opening either to small-diameter tubing for pneumatic signal transmission, or to diaphragm transducers for electrical signal transmission. Only pneumatic pressure probes which are connected in turn by other tubing (usually flexible, and of a larger diameter) to transducers will be considered here. The transducer serves to convert pressure effects to some easily measurable signal such as a mechanical motion or an electrical output. Such transducer signals are then interpreted back in terms of pressure via fundamental relations (as $\Delta p = w\Delta h$ for manometer transducers) or by experimental calibrations (as $\Delta p = K\Delta E$ for diaphragm transducers).

Often, the orifices and associated tubing are subjected to time-varying pressures (e.g., there may be a nonsteady flow, or the probe may be moving, perhaps rotated, through a pressure-gradient field). Thus, while in the steady state, pressure-sensing systems are usually of the nonflow variety, yet in transient, there will usually be a flow in the system accompanied, at the least, by a flow resistance in the orifices and tubings and, in general, by compressible effects within the transducer. As a result, the pressure signal at the transducer lags the impressed pressure at the orifice.

It is the purpose of this paper to indicate a simple expression which may serve to predict the response behavior of a pressure-sensing system when it is subjected to a variety of time-varying pressure signals. The problem will be approached from the fluid-mechanics point of view, rather than from the control viewpoint (via the Laplace-transform natural-frequency damping-factor approach) [1, 2],[1] and rather than from the viewpoint of wave-motion propagation (via acoustical theory) [3, 4].

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, Atlantic City, N. J., November 29–December 4, 1959, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, October 19, 1959. Paper No. 59—A-289.

Elementary Theory

A single time constant can describe the response behavior of only the simplest of systems. Fortunately, however, it is often adequate to distinguish between but two regions in a pressure-sensing system.

A In one of these regions, call it the capillary portion of the system, *all* the fluid resistance is "lumped." (Note: If the system has more than one capillary, an equivalent length may be defined, Ref. [5], page 7, as

$$L_e = L_1 + L_2\left(\frac{d_1}{d_2}\right)^4 + \ldots + L_n\left(\frac{d_1}{d_n}\right)^4 \qquad (1)$$

where d_1 is arbitrarily taken as the reference diameter.)

B In the second region, which will be called the transducer portion of the system, *all* the fluid mass is "stored." (Note: The volume of this transducer portion should include the volume of the equivalent capillary.)

The system, for which a linear expression describing transient behavior is to be developed, is illustrated in Fig. 1.

In the smooth-walled, circular capillary, an axisymmetric, laminar flow of a compressible, viscous fluid takes place in transient. Discussion will be limited to the manometric type of transducer since it represents the more general case (system vol-

[1] Numbers in brackets designate References at end of paper.

Nomenclature

A = flow area of capillary, power of variable
A_e = equivalent area
a_i = $e^{-t/K}$
D = ID of manometer tube
d_n = ID of any capillary
E = emf
e = base of natural logarithm
h = height of manometer fluid
K = lumped constants of system, time constant
L_e = length of equivalent capillary
L_n = length of any capillary
p = capillary pressure
p_e = environment pressure at any time
p_f = final environment pressure
p_i = initial environment pressure
p_r = reference pressure
p_t = transducer pressure at any time
R = gas constant, result
r = any radius in capillary
r_0 = capillary radius from center to wall
T = absolute temperature
t = time
V = system volume at any time, variable
v = axial component of fluid velocity in capillary
w = specific weight of manometer fluid
Δ = finite difference
ϵ = uncertainty
μ = dynamic viscosity coefficient
ρ_f = fluid density at capillary-transducer interface
τ = shear stress
$v_{i,i} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}$

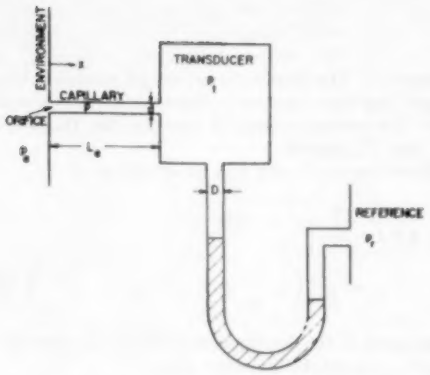


Fig. 1 Pressure-sensing system to be analyzed

ume being a function of time). To further preserve generality, a distinction is made between environment and reference pressures although, experimentally, initial environment pressure would be stepped most conveniently to the reference pressure.

Such a system may be characterized [6] by writing a flow balance at the capillary-transducer interface (i.e., at $z = L_c$).

Mass flow rate into transducer = rate of increase of mass in transducer

$$\int_0^{r_0} \rho_i v dA = \rho_i \frac{dV}{dt} + V \frac{d\rho_i}{dt} \tag{2}$$

The flow balance is localized in this manner to simplify the analysis, since at this interface the density term has consistent meaning.

Appropriate expressions will now be developed for the various terms in equation (2).

A Four distinct pressures may be noted in the system.

1 The environment pressure p_e undergoes a step change with time. At $t = 0^-$, $p_e = p_i$; at $t = 0^+$, $p_e = p_f$.

2 The capillary pressure, $p = f(x, r, t)$ may be defined in terms of its variation within the capillary by writing instantaneous force balances on an annular element in both the axial and radial directions.

Axial Direction, Fig. 2(*A*)

$$\text{Net shear force} = -\frac{\partial}{\partial r}(rd\theta dx\tau_{rx})dr \simeq d\theta dx dr \frac{\partial[\tau_{rx}r]}{\partial r} \tag{3}$$

$$\text{Net pressure force} = -\frac{\partial}{\partial x}(rd\theta drp)dx \simeq d\theta dr dx \frac{\partial[rp]}{\partial x} \tag{4}$$

Assumption: Second-order terms are negligible.

Now, if the ΣF_x are set equal to zero

Assumption: Inertia forces are negligible compared with viscous forces in the axial direction.

there results

$$-\frac{\partial p}{\partial x} = \frac{1}{r}\frac{\partial}{\partial r}(\tau_{rx}r) \tag{5}$$

Radial Direction, Fig. 2(*B*)

$$\text{Net shear force} = -(rd\theta dr)\frac{\partial \tau_{xr}}{\partial x}dx \tag{6}$$

$$\text{Net pressure force} = -(rd\theta dx)\frac{\partial p}{\partial r}dr \tag{7}$$

Assumption: Second-order terms are negligible.

Now, if the ΣF_r are set equal to zero

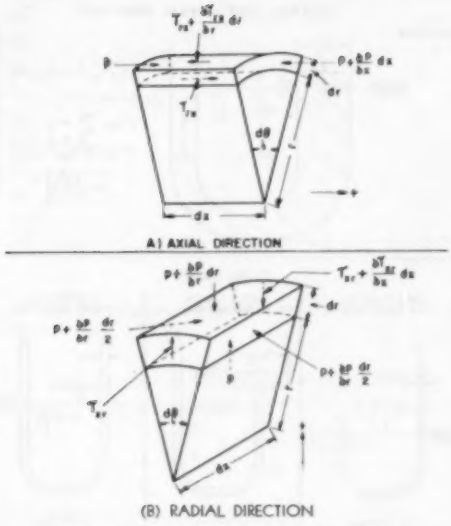


Fig. 2 Various stresses acting on the fluid element

Assumption: Inertia forces are negligible compared with viscous forces in the radial direction.

there results

$$-\frac{\partial p}{\partial r} = \frac{\partial \tau_{xr}}{\partial x} \tag{8}$$

NOTE: Equations (5) and (8), which indicate axial and radial equilibrium in the absence of inertia and body forces, may be recognized as forms of the Navier-Stokes Equations in cylindrical co-ordinates with axial symmetry assumed, [7] p. 49, [8] p. 37. Also, by taking moments about any axis, under equilibrium conditions, it may be seen, [7] page 41, [8] page 20, that pairs of shearing stresses with subscripts which differ only in their order are equal, i.e., $\tau_{rx} = \tau_{xr}$

3 The transducer pressure, $p_i = f(t)$, is the function desired. Initial and final conditions are as follows:

$$\text{at } t = 0, \qquad p_i = p_i$$

$$\text{at } t = \infty, \qquad p_i = p_f$$

NOTE: Transducer dimensions are such that the pressure p_i may be considered uniform throughout at any time.

4 The reference pressure p_r is a constant and needs no further definition. It may be greater than, equal to, or less than the transducer pressure.

B The density of the fluid at the capillary-transducer interface may be defined as

$$\rho_i = \frac{p_i}{RT} = f(t) \tag{9}$$

Assumption: Relations between thermodynamic quantities of fluid are closely approximated by the equation of state of an idealized gas.

The rate of change of density in the transducer is

$$\frac{d\rho_i}{dt} = \frac{1}{RT}\frac{dp_i}{dt} \tag{10}$$

Assumption: Fluid density changes isothermally in transducer.

C Lumping the volume of the system in the transducer, just as resistance of the system was lumped in the capillary, V_∞, is assigned to designate the steady-state volume of the entire sys-

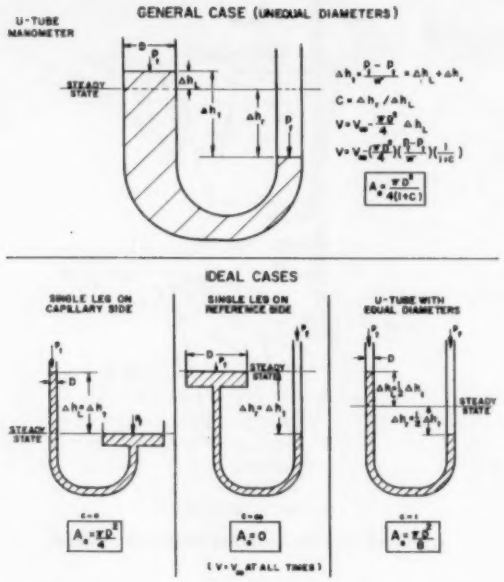


Fig. 3 Definitions of equivalent area for various manometers

tem, from orifice to surface of the manometric fluid. The system volume at any time is then

$$V = V_\infty - \frac{A_e}{w}(p_f - p_i) \tag{11}$$

where A_e is an equivalent area dependent on the geometry of the manometer used (see Fig. 3).

The rate of change of volume is

$$\frac{dV}{dt} = \frac{A_e}{w}\frac{dp_i}{dt} \tag{12}$$

D The instantaneous axial component of the velocity in the capillary is obtained by separating variables and twice integrating equation (5), the expression for the axial pressure gradient in the capillary, where, by definition

$$\tau_{zr} = -\mu\frac{\partial v}{\partial r} \tag{13}$$

and where, as a direct consequence of equations (8) and (17), $\partial p/\partial x \neq f(r)$, as indicated in greater detail at a more appropriate place in the development.

There results

$$-\frac{1}{4\mu}\frac{\partial p}{\partial x}r^2 = -v + C_1 \ln r + C_2 \tag{14}$$

Introducing boundary conditions and evaluating

$$v = \text{finite at } r = 0, \qquad \therefore C_1 = 0$$

$$v = \text{zero at } r = r_0, \quad \therefore C_2 = -\frac{1}{4\mu}\frac{\partial p}{\partial x}r_0^2$$

Assumption: There is no slippage of fluid at the solid boundaries, Ref. [9] p. 168.

Thus an expression for the instantaneous axial velocity at any radius becomes

$$v = -\frac{1}{4\mu}\frac{\partial p}{\partial x}(r_0^2 - r^2) \tag{15}$$

Assumption: The time constant of the system is large compared with the time required for the establishment of a parabolic velocity distribution typical of flow in the Hagen-Poiseuille regime, Ref. [7] page 66.

The flow balance (2), may now be written as

$$-\int_0^{r_0}\frac{p_i}{RT}\frac{1}{4\mu}\frac{\partial p}{\partial x}(r_0^2 - r^2)2\pi\, r dr = \frac{1}{RT}\left(p_i\frac{A_e}{w} + V_\infty - (p_f - p_i)\frac{A_e}{w}\right)\frac{dp_i}{dt} \tag{16}$$

At this point in the development it becomes convenient to impose on the solution the limitation that

$$(p_f - p_i) \ll p_i \tag{17}$$

NOTE: This limitation does not imply the assumption of an incompressible fluid within the system, but only that the fluid is essentially "uncompressed" within the capillary. Indeed, it is just because the fluid within the transducer is compressible (i.e., fluid density changes as transducer pressure changes with time) that there exists a time delay. Inertia of the manometric fluid would likewise contribute to the time lag [10] page 2, but all inertia effects have been neglected in this analysis. By limiting the analysis to those cases where pressure changes are small compared with pressure level, nonlinearities,[2] which must arise [11] when density is considered to be a function of pressure, are avoided.

Nonlinearities introduced by end effects (orifices) or flow rectification (different flow coefficients in and out of the capillary) have also been neglected [12] page 102.

Assumption: Nonlinear pressure effects are negligible compared with the linear pressure drop across the capillary.

Equation (16) may be reduced to a first-order, first-degree, nonlinear differential equation, and as such precludes definition of a time constant; but a time constant is precisely what is being sought. Thus seeking a linearized form of equation (16), is just the reason for the introduction of equation (17). For, concerning pressure level, equation (17) implies that p_v, which is bounded by p_i and p_f, may be well approximated by its average value, i.e., by

$$\bar{p}_v = \frac{\int_0^t p_i dt}{t} \tag{18}$$

But, concerning pressure variations with time, dp_i/dt cannot be replaced by $d\bar{p}_i/dt$ (which latter of course equals zero), since system lag derives from the very fact that $p_i = f(t)$.

NOTE: Mathematically, equation (17) indicates that (concerning level) p_i may be approximated by either p_i or p_f; but practically a departure will be made from the limitations of equation (17) ($\ll$ is really undefined anyway); and thus $\bar{p}i$ is a more realistic approximation to p_i (as will be seen in the section on comparisons).

Making use of the approximation, $p_i \approx \bar{p}_v$, the flow balance as given by equation (16) simplifies to

$$-\left(\frac{\pi d_i^4 \bar{p}_i}{128\mu}\right)\frac{\partial p}{\partial x} = \left[\frac{A_e}{w}(2p_i - p_f) + V_\infty\right]\frac{dp_i}{dt} \tag{19}$$

Here it is observed that the imposition of equation (17) on the solution is tantamount to staring that $\partial p/\partial x$ is a constant with respect to x, and that $p \neq f(r)$. For note that, when the fluid is

[2] Nonlinearity here means that effects are not proportional to causes. That is, there is no simple straight-line relationship between input and output, no matter how the plotting scales are distorted.

considered "uncompressed" within the capillary, and when the fluid flows laminarly in the axial direction (i.e., $v_r = v_\theta = 0$), then continuity (i.e., $v_{i,i} = 0$) insures that $\partial v/\partial x = 0$. Thus, $v \neq f(x)$, and by substituting equation (13) successively in equations (5) and (8) then

$$\frac{\partial p}{\partial x} \neq f(x)$$

and

$$\frac{\partial p}{\partial r} = 0$$

Consequently, by separating variables, equation (19) may be written as

$$-\left(\frac{\pi d_i^4 \bar{p}_i}{128\mu}\right)\int_{p_f}^{p_i} dp = \left[\frac{A_s}{w}(2\bar{p}_i - p_f) + V_\infty\right]\frac{dp_i}{dt}\int_0^{L_s} dx \quad (20)$$

Assumption: Fully developed laminar flow is established between points of integration, Ref. [7] page 149.

Upon integration the first-order, linear expression

$$\frac{dp_i}{dt} = \frac{1}{K}(p_f - p_i) \quad (21)$$

is obtained where K lumps the constants of the system, and is defined mathematically as

$$K = \frac{128\mu L_s}{\pi d_i^4 \bar{p}_i}\left[\frac{A_s}{w}(2\bar{p}_i - p_f) + V_\infty\right] \quad (22)$$

K will now be shown to be the time constant of the pressure-sensing system. Equation (21) has the general solution [13]

$$p_i = Ce^{-t/K} + \frac{1}{K}e^{-t/K}\int_0^t p_f e^{t/K} dt \quad (23)$$

For the case of a pressure-sensing system subjected to a step-change in environment pressure, the initial condition is

$$\text{at } t = 0, \quad p_i = p_{i_1}, \quad \therefore C = p_{i_1}$$

and at any time, p_f = constant. The general solution reduces to[3]

$$(p_f - p_i) = (p_f - p_{i_1})e^{-t/K} \quad (24)$$

Thus K indicates the time required (after the step change) for the environment-transducer pressure difference to be reduced to 1/eth of the initial difference. In other words, K is the time required for the pressure-sensing system to indicate a pressure corresponding to approximately 63.2 per cent of the initial pressure difference. This is, of course, the meaning reserved for the time constant of a first-order linear system so that K equals the time constant. Caution: For higher order, variously damped systems, the time constant has no such significance, [2], page 8. That is, if the damping factor of a second-order system was very small, one would expect a decaying oscillatory response, and the 63 per cent recovery would have little meaning; conversely, if the damping factor was ≥ 1, the second-order response would appear much the same as a first-order response, and the 63 per cent recovery would appear to have significance. For Reynolds numbers below 2000, a capillary acts like a highly damped system and therefore appears much the same as a first-order, linear system.

[3] A similar expression is given by H. L. Weissberg in Ref. [14], however, his K is defined differently.

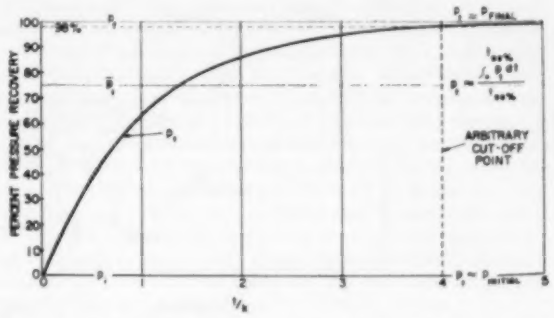


Fig. 4 Several approximations to p_i for a step change in environment pressure

The time required to reduce the pressure difference to any acceptable value may be predicted from the expression

$$a_t = \frac{p_f - p_i}{p_f - p_{i_1}} = e^{-t/K} \quad (25)$$

which tabulates as

a_t	0.5	0.368	0.2	0.1	0.005	0.01
t	0.7K	K	1.6K	2.3K	3K	4.6K

It is now possible to evaluate $\bar{p}_i$ and thus to obtain explicit expressions for both the time constant (K) and the recovery time (t).

Inserting p_i from equation (24) and t from equation (25) into equation (18), then

$$\bar{p}_i = p_f - (p_f - p_{i_1})\left(\frac{1 - a_t}{\ln \dfrac{1}{a_t}}\right) \quad (26)$$

Now, since p_i asymptotically approaches but can never equal p_f in any practical time, a cut-off point for a_t must arbitrarily be decided upon if a realistic $\bar{p}_i$ is to be evaluated (see Fig. 4). For simplicity, an a_t corresponding to 98 per cent recovery was chosen. Under this condition

$$\bar{p}_i = p_f - \frac{(p_f - p_{i_1})}{4} \quad (27)$$

where $\bar{p}_i$ represents the average value of p_i from time zero to a time corresponding to the 98 per cent recovery point, i.e., during the greater portion of the transient phenomenon.

The time constant now becomes

$$K = \frac{128\mu L_s}{\pi d_i^4 \bar{p}_i}\left[\frac{A_s}{w}\left(\frac{p_f + p_{i_1}}{2}\right) + V_\infty\right] \quad (28)$$

and the general expression for recovery time becomes

$$t = \frac{128\mu L_s}{\pi d_i^4 \bar{p}_i}\left[\frac{A_s}{w}\frac{(p_f + p_{i_1})}{2} + V_\infty\right]\ln\left(\frac{p_f - p_{i_1}}{p_f - p_i}\right) \quad (29)$$

Experiment

An experiment was designed to allow determination of the transient behavior for a variety of pressure-sensing systems. The over-all test setup is illustrated in Fig. 5. Pressure levels of 1/4, 1/2, 1, and 2 atmospheres were used in this experiment. Cellophane was chosen as the diaphragm material, with the understanding that no appreciable energy was imparted to the fluid by the bursting diaphragm, and with the finding that it was econom-

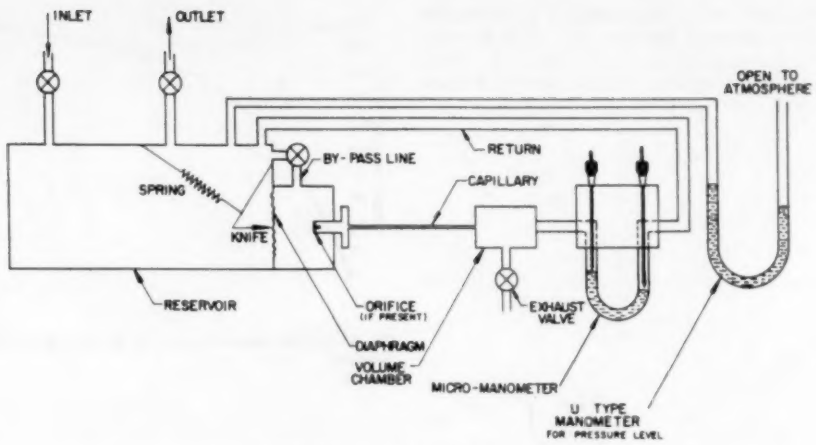


Fig. 5 Schematic of experimental set-up

ical, safe, and easy to manufacture in the form of gaskets. An Exacto knife was electrically actuated by a solenoid to puncture the diaphragm. Ref. [15] describes an experiment in which glass diaphragms were used. Ref. [16], page 697, describes a "high-speed" orifice design which was also tested. The micromanometer was specially designed to allow electrical indication of the manometer-fluid surface location. Ref. [17], page 3, mentions use of a similar system. The response time was obtained auto matically by an electric timer which started counting when the knife burst the diaphragm, and stopped when the moving mercury column in the micromanometer contacted the micrometer rod (preset for the desired per cent recovery).

Experimental Uncertainty

The important question as to how the uncertainties of individual measurements are propagated into the results will be considered now.

Following Ref. [18], page 3, use will be made of a second-power form of the logarithmically differentiated expression relating variables and result; i.e.,

$$\frac{\epsilon_R}{R} = \pm\left[\left(A_1\frac{\epsilon_1}{V_1}\right)^2 + \left(A_2\frac{\epsilon_2}{V_2}\right)^2 + \ldots + \left(A_n\frac{\epsilon_n}{V_n}\right)^2\right]^{1/2} \tag{30}$$

where

ϵ = uncertainty, or possible value of error
A_1, A_2, A_n = powers to which respective variables occur in expression relating variables and result
R = result
V_1, V_2, V_n = variables

Thus the system will be analyzed using observed uncertainties in individual measurements to form conclusions as to the worth of experimental answers. Considering equation (29), it is noted that, in general, the term formed with V_∞ far outweighs in importance the other two terms in the expression. So, in the uncertainty considerations, this term only will be considered. Furthermore, the analysis will be limited to the $^1/_4$-atmosphere pressure level since the greatest experimental uncertainties are encountered there.

All variables are first listed in order of importance, and then with the help of equation (30) several uncertainties to be expected in the experimental determinations of time are evaluated. The relative uncertainty in time (ϵ_t/t) is found to be approximately ± 6 per cent for any per cent recovery. However, for tests on a given geometry, the length, volume, and diameter are fixed quantities and the experimental scatter should be much less (say ± 3 per cent) than that predicted by the foregoing analysis which concerns uncertainties in measurement level as well.

Comparisons

Graphical comparisons were made between IBM-704-computer solutions to the following equations.

$$t_1 = B/p_i\left[V_\infty + \frac{A_s}{w}\frac{(p_i + p_f)}{2}\right]\ln\frac{1}{a_i}$$

This is the linear equation developed in this paper, and is based on $p_i \approx p_f - (p_f - p_i)/4$.

$$t_2 = B/p_f\left[V_\infty + \frac{A_s}{w}(p_f)\right]\ln\frac{1}{a_i}$$

This is a linear equation based on $p_i \approx p_f$.

$$t_3 = B/p_i\left[V_\infty + \frac{A_s}{w}(2p_i - p_f)\right]\ln\frac{1}{a_i}$$

This is a linear equation based on $p_i \approx p_i$.

$$t_4 = B/p_f\left[V_\infty \ln\frac{1}{a_i b_i} + \frac{A_s}{w}\left(p_f \ln\frac{1}{a_i} + 3p_f \ln b_i\right)\right]$$

This is a nonlinear equation based on Equation (3) of [5], page 6.

$$t_5 = B/p_f\left[\bar{V}\ln\frac{1}{a_i b_i} + \frac{A_s}{w}\left(2p_f \ln\frac{b_i}{a_i} - p_i \ln\frac{1}{a_i b_i}\right)\right]$$

This is a nonlinear equation based on Equation (8) of [19], page 14.

where:

$B = 128\mu L_1/\pi d_1^4 F$
$F = 1/[(L_2/L_1)(d_1/d_2)^4 + 1]$
$V_\infty = V_1 + V_2 + V_m$
$\bar{V} = \frac{KV_1}{2} + \frac{(K+1)V_2}{2} + V_m$
V_1 = volume of first capillary
V_2 = volume of second capillary

V_m = volume of transducer
$a_i = (p_f - p_i)/(p_f - p_i)$
$b_i = (p_f + p_i)/(p_f - p_i)$

Experimental data obtained in the present work were superimposed on certain of these computer solutions, together with the experimental uncertainty envelope, to indicate whether deviations between the various solutions were significant in view of experimental uncertainties.

The several cases considered include:

1 Usual capillary-transducer system in which all resistance is considered in the capillary and all volume is considered in the transducer.

2 Orifice-capillary-transducer system in which a definite nonlinear pressure drop occurs at orifice.

3 Capillary-capillary-transducer system in which the resistance and volume are both distributed throughout the system.

Conclusions

All conclusions are based on the graphs given in Figs. 6 through 9 and others not illustrated (note that results in Figs. 6, 7, and 9 are plotted on a semilog co-ordinate system to bring out the linearity of the proposed equation).

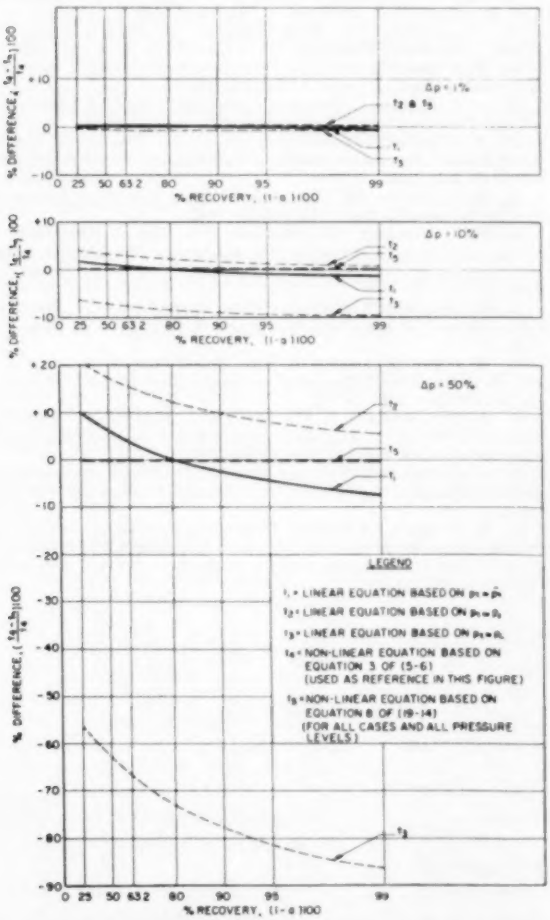

Δp = 1%
Δp = 10%
Δp = 50%
% RECOVERY, (1 - a) 100
LEGEND
NON-LINEAR EQUATION BASED ON EQUATION 3 OF (5-6) (USED AS REFERENCE IN THIS FIGURE)
NON-LINEAR EQUATION BASED ON EQUATION 8 OF (19-14) (FOR ALL CASES AND ALL PRESSURE LEVELS)

Fig. 6 Comparison of several expressions for response time after various step changes in environment pressure

1 From Fig. 6 it is concluded that all equations, linear and nonlinear, reliably predict response times for pressure differences on the order of 1 per cent of the pressure level. It is further concluded from this figure that the nonlinear equations are less affected by increases in the size of the step change in pressure than are the linear equations. Finally, it is noted that the linear equation developed in this paper, namely, equation (29), is in closest agreement with the nonlinear equations over the transient range of most general interest (i.e., from 25 to 99 per cent recovery), no matter what the size of the step change in pressure.

2 The advantage of the linear over the nonlinear equations is that the constant slopes of the linear plots insure that time constants can be defined, while definition of a time constant is precluded by the varying slope of any nonlinear equation.

3 From Fig. 7, the excellent agreement at various pressure levels between experimental points and the linear equation proposed in this paper is noted.

4 In Fig. 8, the effect of change in length and change in diameter of the capillary on this time constant of the linearized

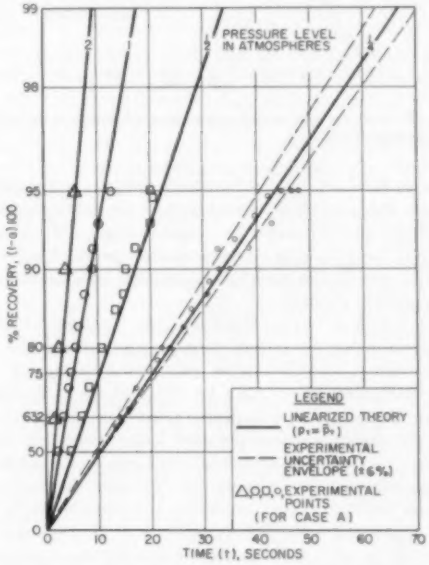


Fig. 7 Comparison of experiment versus linearized theory at various pressure levels

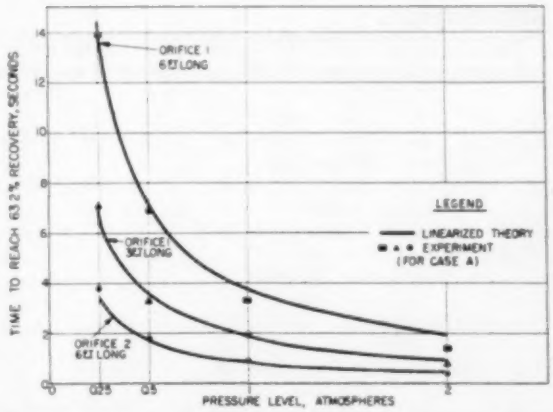


Fig. 8 Effect of change in length and change in diameter on the time constant of the linearized expression

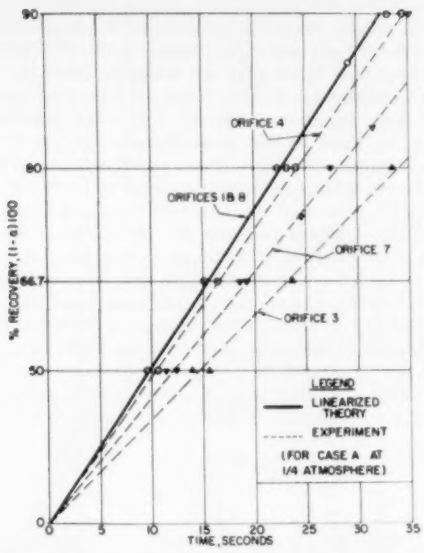


Fig. 9 Effect of various orifice geometries on the time constant of the linearized expression

expression is noted. Experimental points are superimposed on this plot. Response time is seen to vary directly with length, i.e., halving the length halves the response time. Much more important is the reduction in response time realized by increasing the capillary diameter since time varies inversely as the 4th power of the capillary diameter.

5 In Fig. 9, it is confirmed that the "high-speed" orifice of Ref. [16] is faster than usual long-passage orifices and that, in general, orifices of small lengths compared with capillary lengths and of diameters greater than $^1/_4$ of the capillary diameters introduce nonlinear pressure drops which are insignificant compared with the quasi-linear pressure drops across the capillaries.

6 In view of the experimental evidence, and considering the good agreement between equation (29) and the more general (although more unwieldy) nonlinear equations of Refs. [5] and [19], it is recommended that equation (29) be used to represent transient response time for pressure-sensing systems because of its simplicity, utility, and for the ease with which it may be related to the elementary notions of fluid mechanics.

Acknowledgments

The valuable assistance of W. G. Steltz who set up the Fortran source program used in the IBM-704-computer comparisons of the various expressions for response time, of M. H. Wachs who carried out some of the experimental and analytical work, of G. M. Luongo who designed most of the experimental apparatus, of J. Fieldhouse who constructed most of the experimental apparatus, of D. Foellner who constructed the control box, and of A. M. Zampitella who helped throughout the experimental portion of this work are gratefully acknowledged by the author.

References

1 C. B. Schuder and R. C. Binder, "The Response of Pneumatic Transmission Lines to Step Inputs," JOURNAL OF BASIC ENGINEERING, TRANS. ASME, series D, vol. 81, 1959, pp. 578–584.

2 E. R. Quandt, "An Analysis of Pressure Transducer Transient Response," Westinghouse Atomic Power Division Report WAPD-TH-448, September, 1958.

3 G. J. Delio, G. V. Schwent, and R. S. Cesaro, "Transient Behavior of Lumped-Constant Systems for Sensing Gas Pressures," NACA TN 1988, December, 1949.

4 I. Taback, "The Response of Pressure-Measuring Systems to Oscillating Pressures," NACA TN 1819.

5 A. R. Sinclair and A. W. Robbins, "A Method for the Determination of the Time Lag in Pressure Measuring Systems Incorporating Capillaries," NACA TN 2793, September, 1952.

6 R. P. Benedict, "The Response of a Pressure-Sensing System," Westinghouse Steam Division, Report EM-339, September, 1958.

7 H. Schlichting, "Boundary Layer Theory," McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

8 S. I. Pai, "Viscous Flow Theory," D. Van Nostrand Company, Inc., Princeton, N. J., 1956, vol. 1, "Laminar Flow."

9 Sir J. Jeans, "An Introduction to the Kinetic Theory of Gases," Cambridge University Press, Cambridge, England, 1952.

10 F. A. MacMillan, "Liquid Manometers With High Sensitivity and Small Time-Lag," NACA N-30309, 1941.

11 R. C. Johnson, "Averaging of Periodic Pressure Pulsations by a Total-Pressure Probe," NACA TN-3568, October, 1955.

12 A. S. Iberall, "Attenuation of Oscillating Pressures in Instrument Lines," *Journal of Research*, National Bureau of Standards, RP-2115, July, 1950.

13 D. R. Harper, III, "Thermometric Lag," National Bureau of Standards Scientific Paper 185, March, 1912.

14 H. L. Weissberg, "The Response Time of Small Pitot Tubes," Union Carbide and Carbon Corporation Report KLI-2616, October, 1953; available from the Office of Technical Services as PB-AECU-3242.

15 R. Masnovi, "Transducer System Response Tests," Westinghouse Atomic Power Division Report WAPD-TH-450, October, 1958.

16 W. S. Bradfield and G. E. Yale, "Small Tubes With Fast Pressure Response Time," *Journal of the Aeronautical Sciences—Reader's Forum*, October, 1951.

17 J. F. Swindells, J. R. Coe, Jr., and T. B. Godfrey, "Absolute Viscosity of Water at 20 C," *Journal of Research*, National Bureau of Standards, vol. 48, January, 1952, RP-2279.

18 S. J. Kline and F. A. McClintock, "Describing Uncertainties in Single-Sample Experiments," *Mechanical Engineering*, vol. 75, 1953, pp. 3–8.

19 R. C. Bauer, "A Method of Calculating the Response Time of Pressure Measuring Systems," Arnold Engineering Development Center TR-56-7, November, 1956; available from Office of Technical Services as PB-131321.

Dynamics of Heat Exchangers and Their Models

HERMAN THAL-LARSEN
Associate Professor of Mechanical Engineering, University of California, Berkeley, Calif. Mem. ASME

The author rejects, momentarily, actual experimentation with real heat exchangers for logical cogitation and arm-chair experiments. Observed as mental images or models, these heat exchangers reveal their basic characteristics—characteristics that will be partially obscured in an actual experiment by secondary effects inherent in the equipment. These basic characteristics, once perceived, may be used to correlate by means of frequency-response diagrams many results heretofore diverse and seemingly unrelated, that have been reported in the literature. Included among such results are not only the dynamics of various types of heat exchangers but also the thermal dynamic interaction between fluids and the confining pipe used for their transport. In addition, the author warns that all too frequently the use of the average temperature in heat-exchanger dynamics is based on faulty logic.

TRANSFER of heat between fluids through metal walls and its transport down metallic tubes is an ever-recurring problem in industry. For many years this problem has been solved, giving attention only to the steady-state part of the solution. The transient part of the solution has only recently become of sufficient importance to draw considerable effort in research. This shift toward more emphasis upon transient behavior has been caused by two factors; namely, the greater application of automatic control which is absolutely dependent upon the transient state, and the greater speeds at which some heat processes now operate, thus demanding prior knowledge of transient behavior.

Mathematical analysis of the transient state is much more complex than the relatively simple analysis required for the steady state. Differential equations for the transient state require, in addition to the parameters associated with steady-state behavior, a new set of parameters connected with dynamic behavior. These more complex equations can still be solved in terms of steady-state solutions. But the solutions are of a unique nature. They are sinusoids and neither differentiation nor integration is able to alter their shape. Joseph Fourier showed us more than a hundred years ago that such sinusoidal solutions and periodic transients are mutually convertible. It is interesting to note that his researches in this connection were also concerned with transient heat transfer.

The frequency-response method of analysis is today the most commonly used method for studying the transient behavior of heat exchangers. This is because frequency-response data are compatible, while transient-response data are not. The over-all dynamic behavior of several pieces of equipment can readily be synthesized if their frequency-response characteristics are available. This cannot be done as conveniently when only transient-response data are given.

We will therefore emphasize the frequency-response aspects of heat-exchangers and pipelines in this paper. We will survey the literature for pertinent data and we will compare the data on a frequency-response basis as this will make the information immediately useful to both the control engineer and to the designer of heat-exchanger equipment.

However, comparisons of the dynamic behavior of apparently dissimilar pieces of equipment must be made on a rational basis, using the most fundamental of the parameters that characterize dynamic behavior. Our problem then is first to discover these basic parameters out of the multitude which can be written,

Contributed by the Instruments and Regulators Division and presented at the Annual Meeting, Atlantic City, N. J., November 29–December 4, 1959, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters August 13, 1959. Paper No. 59—A-117.

Nomenclature

A = heat-transfer area, sq ft
a_1 = parameter, defined by Equation (50), 1/ft
a_2 = parameter, defined by Equation (51), dl
C_o = capacitance of liquid slug in tube, Btu/deg F
C_s = capacitance of liquid in shell, defined by Equation (25), Btu/deg F-ft
c = specific heat, Btu/lb-deg F
D = ratio, defined by Equation (58), ft
d, d_1 = inside tube diameter, ft
d_2 = outside tube diameter, ft
e = base of natural logarithms, 2.718 . . . , dl
h = heat-transfer film coefficient, Btu/sec-sq ft-deg F
j = $\sqrt{-1}$
K_1 = constant, defined by Equation (5), dl
K_2 = constant, appearing in Fig. 11, deg F
K_m = parameter, defined by Equation (53), 1/sec
K_t = parameter, defined by Equation (52), 1/sec
k = constant in Equation (36), dl
L = length, ft
M = mass, lb
n = an integer, 0, 1, 2, 3, . . . , dl
T_{s1}, T_{s2} = steady average temperatures, deg F
T_1, T_2 = temperature perturbations, deg F
t = time, sec
t_1 = Equation (27), sec
t_2 = Equation (30), sec
Δt = pulse duration, defined by Equation (31), sec
U = over-all coefficient of heat transfer, Btu/sec-sq ft-deg F
V_0 = steady average flow velocity, ft/sec
V = flow velocity perturbation, ft/sec
W_0 = steady average weight flow rate, lb/sec
W = weight-flow-rate perturbation, lb/sec

(Continued on next page)

depending upon the various permutations and combinations allowed, with the design and operating variables of our equipment. Quoting from "How to Hunt a Submarine," by Morse and Kimball [1][1]: "It is important first to obtain an over-all quantitative picture of the operation under study. One must first see what is similar in operations of a given kind before it will be worth-while seeing how they differ from each other. In order to make a start in so complex a subject, one must ruthlessly strip away details (which can be taken into account later), and arrive at a few broad, very approximate 'constants of the operation.' By studying the variations of these constants, one can then perhaps begin to see how to improve the operation. It is well to emphasize that these constants which measure the operation are useful even though they are extremely approximate . . ."

We will discover these constants of the operation by the simplest of analyses which are promised in advance to be neither rigorous nor difficult, but nevertheless illuminating. Then we will test these constants against data reported in the literature. Such data were usually obtained at the expense of considerable effort and they should be consulted by the reader wishing greater detail in the analysis. Data which appear in the form of time or space transients [2] or as polar plots [3, 4] in the literature have been converted to frequency-response diagrams on Cartesian co-ordinates with logarithmic scales for the purpose of this paper.

Because we wish to emphasize the dynamic aspects of heat-exchange processes we shall ignore the static relationships by normalizing the amplitude ratio of our frequency-response curves; i.e., we shall only show the dynamic amplitude ratio relative to the static or very slow dynamic amplitude ratio as unity. The static relationships are today rather well understood and need no discussion here.

Finally, we shall show that the commonly used [5, 6] average temperature in the analysis of heat-exchanger dynamics is all too often based upon faulty logic—that its use can lead to erroneous conclusions concerning control-system stability.

Model 1

Fig. 1 represents a simple heat exchanger. The water in contact with the steam chest is well stirred assuring a uniform temperature throughout mass M. Temperatures T_1, T_2, and T_3 represent fluctuations from their mean values T_{01}, T_{02}, and T_{03}. Higher subscript numbers indicate higher mean temperatures.

[1] Numbers in brackets designate References at end of paper.

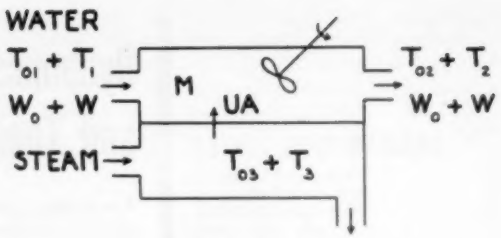


Fig. 1 Model 1

W is a weight-flow perturbation about W_0. Let c represent the specific heat of water. U is the over-all coefficient of heat transfer between the steam and the water above it through the thin metal plate of area A. Other simplifying assumptions associated with the model are implied in the equations which follow. The frequency response relating sinusoidal fluctuations in T_1 to those in T_2 is given by

$$\frac{T_2(j\omega)}{T_1(j\omega)} = \frac{1}{1 + j\omega\tau_{da}} = \frac{1}{1 + j\Omega_{da}} \tag{1}$$

Equation (2) defines τ_{da}. It is the average time required

$$\tau_{da} = \frac{M}{W_0} \tag{2}$$

by a particle of water to traverse the heat exchanger, hence it is also known as the average dwell time. And Ω_{da} is the dimensionless circular frequency normalized on τ_{da} as defined by

$$\Omega_{da} = \tau_{da}\omega \tag{3}$$

The frequency response relating T_3 and T_2 is given by Equation (4),

$$\frac{T_2(j\omega)}{T_3(j\omega)} = \frac{K_1}{1 + j\omega\tau_1} \tag{4}$$

with zero-frequency amplitude ratio K_1 defined by

$$K_1 = \frac{UA}{W_0c + UA} \tag{5}$$

and time constant τ_1 defined by

Nomenclature

δ = equivalent flow diameter of shell-side fluid, ft
π = 3.14159 . . .
ρ = weight density, lb/cu ft
Σ = "sum of"
τ_0 = equivalent dead time, defined by Equation (64), sec
τ_1 = time constant, defined by Equation (6); also by Equation (59); also by Equation (62), sec
τ_2 = time constant, defined by Equation (60); also by Equation (63), sec
τ_3 = time constant, defined by Equation (61), sec
τ_4 = time constant, defined by Equation (48), sec
τ_d = dwell time, defined by Equation (12), sec
τ_{da} = average dwell time, defined by Equation (2), sec
τ_t = time constant for fluid inside tube, defined by Equations (20) and (52), sec
τ_s = time constant for shell-side fluid, defined by Equation (21), sec
τ_{dt} = dwell time for tube fluid of 2-pass heat exchanger, defined by Equation (22), sec
τ_{ds} = dwell time for shell fluid of 2-pass heat exchanger, defined by Equation (23), sec
τ_m = time constant for metal of tube, defined by Equation (53), sec
ω = circular frequency, radians/sec
Ω = dimensionless circular frequency
Ω_1 = dimensionless circular frequency, defined by Equation (8)
Ω_d = dimensionless circular frequency, defined by Equation (15) or (57)
Ω_{da} = dimensionless circular frequency, defined by Equation (3)
Ω_t = dimensionless circular frequency, defined by Equation (38)

Subscripts

t = tube
m = metal

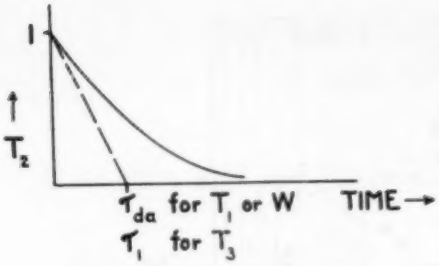


Fig. 2 Impulse response of Model 1

$$\tau_1 = \frac{Mc}{W_0c + UA} \tag{6}$$

We normalize Equation (4) and write it as

$$\frac{T_2(j\omega)}{T_3(jw)} = \frac{1}{1 + j\Omega_1} \tag{7}$$

with Ω_1 the dimensionless circular frequency normalized on τ_1 as defined by

$$\Omega_1 = \tau_1\omega \tag{8}$$

Equations (1), (3), (7), and (8) allow us to describe the dynamic behavior of this heat exchanger, with either T_1 or T_3 as input, and T_2 as output, by means of a single set of frequency-response curves defined by the following:

$$\text{AR} = \frac{1}{(1 + \Omega^2)^{1/2}} \tag{9}$$

$$\text{Phase angle} = -\text{arc tan } \Omega \tag{10}$$

where AR stands for normalized amplitude ratio, and Ω is circular frequency normalized on the appropriate time constant.

A flow perturbation W corresponds to parametric forcing because we effectively perturb the average dwell time of the system τ_{da}. However, small flow oscillations W about the mean value of τ_{da} will not appreciably alter the dynamic picture from that presented by oscillations in temperature T_1.

The frequency response of this heat exchanger may also be found by Fourier transforming the impulse response. For impulses applied at T_1, W, or T_3 the response of T_2 is shown in Fig. 2.

A normalized initial response magnitude is again used. Equations (9) and (10) may also be derived from the Fourier transform of the impulse response shown in Fig. 2.

Model 2

Fig. 3 is the second model. Water flows with velocity V_0 through a thin metal tube of length L. Steam surrounding the tube keeps the metal at a uniform temperature throughout its length. The water flowing within the tube has a steady average temperature gradient in the direction of flow. Assume that each small slug of water inside the tube is perfectly mixed transversely, but that no mixing takes place longitudinally.

We wish to relate sinusoidal fluctuations in steam temperature T_3 to outlet-water temperature fluctuations T_2. Analysis of the problem is complicated by the fact that we must now concern ourselves with both space and time co-ordinates to describe the temperature of the water inside the tube. Hence, partial differential equations are called for. This is the normal approach.

But let us simplify the problem of understanding the basic dynamics of this heat exchanger. What would happen if we applied steam-temperature disturbance T_3 in the form of an impulse? Fig. 4 shows the idealized response of the outlet temperature change T_2. Fourier transforming this response, or, alternatively, Laplace transforming it and subsequently substituting $j\omega$, we get the basic frequency-response characteristic of this heater relating T_2 and T_3,

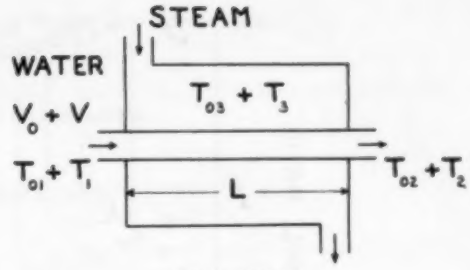


Fig. 3 Model 2

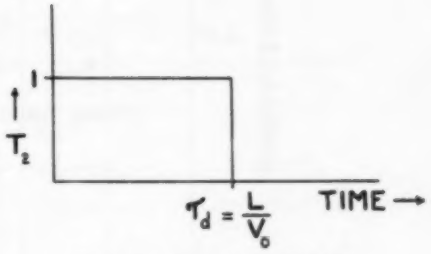


Fig. 4 Impulse response of Model 2

$$\frac{T_2(j\omega)}{T_3(j\omega)} = \frac{1}{j\omega}\left[1 - e^{-j\omega\tau_d}\right] \tag{11}$$

The dwell time τ_d of the liquid within the heat exchanger is given by

$$\tau_d = \frac{L}{V_0} \tag{12}$$

It is more convenient, for plotting purposes, to reduce Equation (11) to the expressions,

$$\text{AR} = \frac{\sin\dfrac{\omega\tau_d}{2}}{\dfrac{\omega\tau_d}{2}} \tag{13}$$

$$\text{Phase angle} = -\frac{\omega\tau_d}{2} \tag{14}$$

Equation (13) has been normalized to zero-frequency amplitude ratio of unity. It appears then that half the dwell time is the prime parameter relating T_2 and T_3 for this type of heat exchanger. We therefore normalize the circular frequency on $\tau_d/2$ as defined by

$$\Omega_d = \frac{\tau_d}{2}\omega \tag{15}$$

The normalized amplitude ratio and angle of lag relating steam-temperature changes T_3 to heated-water fluctuations T_2 are given by

$$\text{AR} = \frac{\sin\Omega_d}{\Omega_d} \tag{16}$$

$$\text{Phase angle} = -\Omega_d \tag{17}$$

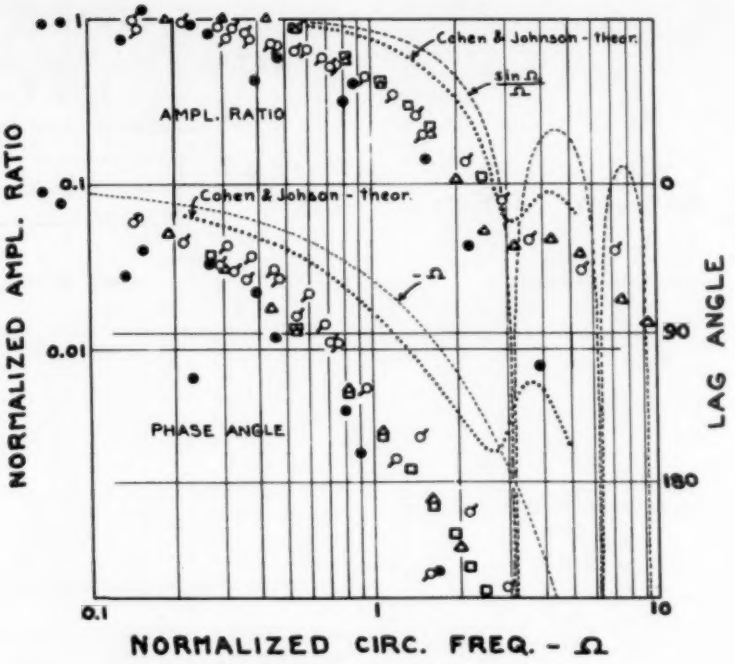


Fig. 5 Frequency response of heat exchangers

Catheron, Goodhue, and Hansen: Steam in shell (steel), water in tubes (copper), 6-passes; 4 × 7 and 2 × 6 tubes, 0.75 in. OD × 0.042-in. wall, heat-transfer area 31 sq ft, water path 24 ft long, 30 lb water in tubes and head, input was sinusoidal stm press changes, output was sinusoidal heated-water temperature changes.
△, 20 gpm, τ_d = 10.8 sec, Ω = (5.4 sec) (ω rad/sec)
□, 40 gpm, τ_d = 5.4 sec, Ω = (2.7 sec) (ω rad/sec)
Lees and Hougen: Steam in shell (steel), water in tubes (copper), 4 passes; 2 bundles of 11 tubes, 0.75 in. OD × 18 BWG, heat-transfer area 34 sq ft, water path 16 ft long, 30 lb water in tubes and head, input was impulselike change in water flow, output was heated-water temperature response. Results were Fourier transformed to obtain freq response. ⊙, 30 gpm, τ_d = 7.2 sec, Ω = (3.6 sec) (ω rad/sec)
Cohen and Johnson: Steam in shell, water in tube, single pass, 1 in. and 2-in. concentric brass pipes, 11.6 ft long, heat-transfer area 4 sq ft, 4.3-lb water in tube.
○, 1.44 lb/sec, τ_d = 3 sec, Ω = (1.5 sec) (ω rad/sec)
Industrial Heat Exchanger: Organic liquid in shell (steel), organic liquid in tubes (steel), 184 tubes 0.75 in. OD × 14 BWG × 192 in. long, heat-transfer area 575 sq ft; input was sinusoidal changes in tube flow, output was sinusoidal changes in tube outlet temperature, heat flow was from tube liquid to shell liquid, frequency response includes unknown and unavoidable thermowell time constant.
●, τ_d = 30 sec, Ω = (15 sec) (ω rad/sec)

Numerical values for Equation (16) are readily available [7]. Both Equations (16) and (17) are shown plotted in Fig. 5. This figure also shows the remarkable frequency-response correlation which exists, through Equation (15), between different heat exchangers operating under various conditions.

Our next concern is the relationship between velocity of flow changes V and the output temperature disturbance T_2. Again, we may apply the concept of the impulse response. A positive impulse V will instantaneously displace the water column inside the tube a small distance to the right. This will have the effect of reducing by the same amount the time which all slugs of water in one tube length have to pick up heat. We assume that the thermal time constant associated with each slug of water is much longer than the dwell time τ_d of the heat exchanger. This is a fair assumption as the exit-water temperature is normally considerably below the steam temperature in the shell space. The impulse response T_2 will then again be as shown in Fig. 4, except for a reversal in sign, a positive impulse producing a temperature drop in T_2. But there will be no sign reversal in case a hot liquid inside the tube is used to transfer heat to a well-mixed cooler liquid in the shell space surrounding the hot tube. Thus flow changes V and tube-liquid exit temperature changes T_2 are related by the frequency-response Equations (13) and (14), keeping in mind to add 180-deg phase lag to Equation (14) if heat transfer is into the tube. The frequency response for tube-flow changes may also be found in Fig. 5.

What effect will temperature changes T_1 have upon T_2? A temperature impulse T_1 will be transmitted down the tube in the form of a hot slug of water with velocity V_0, appearing at the exit after elapse of the dwell time τ_d. We may inject a train of constant-magnitude impulses at T_1, either in rapid sequence or slowly. All pulses will appear at T_2 with the same magnitude,

not necessarily the same magnitude as the injected pulse at T_1. Thus the magnitude of the frequency-response function is a constant, which we normalize to unity. And the phase shift is given by

$$\text{Phase angle} = -\tau_d\omega \tag{18}$$

which, if normalized on half the dwell time, may be written as

$$\text{Phase angle} = -2\Omega_d \tag{19}$$

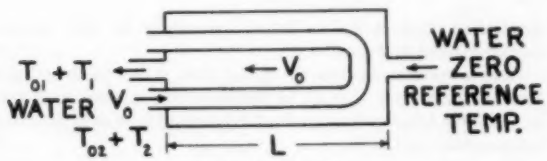


Fig. 6 Model 3

Model 3

Another model, that represented by Fig. 6, presents an interesting problem. Water flows on both sides of the tube which has a very thin metal wall. For simplicity assume that the tube and shell-side velocities are the same V_0. The tube diameter is d, the weight density and specific heat of water ρ and c, respectively. The over-all coefficient of heat transfer through the tube wall is U. The time constant associated with a slug of water inside the tube is given by

$$\tau_t = \frac{\pi d^2\rho c}{4\pi dU} = \frac{d\rho c}{4U} \tag{20}$$

while the time constant for a shell-side slug of water is

$$\tau_s = \frac{\pi\delta^2\rho c}{4\pi dU} = \frac{\delta^2\rho c}{4dU} \tag{21}$$

Here δ is the equivalent flow diameter of the shell fluid.

Dwell time of the liquid inside the tube is given by

$$\tau_{dt} = \frac{2L}{V_0} \tag{22}$$

and dwell time of the shell-side liquid by

$$\tau_{ds} = \frac{L}{V_0} \tag{23}$$

Assume that τ_t is appreciably smaller than $\tau_{dt}/2$.

We wish to find the dynamic relationship between the temperature entering the tube T_2 and that leaving the shell T_1. Again, let us apply an imaginary impulse T_2; i.e., let us suddenly inject a great deal of thermal energy into a small slug of water at the entrance of the tube. Immediately, some of this energy will be transferred through the tube wall into the shell-side slug just emerging from the heat exchanger. Assume that this jump in temperature of the shell-side slug is $T_1 = 100$ deg F, at time $t = 0$. The tube-side slug continuously transfers energy to the shell-side fluid at a rate given by

$$\text{Rate} = \frac{T_2}{\tau_t} C_0 \tag{24}$$

where C_0 is the capacitance of the tube slug. The energy absorbed by the shell-side fluid per unit time is distributed over a shell-side fluid length equal to the relative velocity between the entering tube-side slug and a shell-side slug; namely, a length of shell-side fluid equal to 2 V_0. The capacitance per unit length of the shell-side fluid is

$$C_s = \frac{\pi}{4}\delta^2\rho c \tag{25}$$

We may now calculate the initial temperature pulse by solving Equation (26) for T_2:

$$\frac{\frac{T_2}{\tau_t} C_0}{2V_0C_s} = T_1 = 100 \tag{26}$$

Next, we calculate the temperature of the hot slug inside the tube after a time interval given by

$$t_1 = \frac{L}{V_0} = \tau_{dt}/2 \tag{27}$$

At this time the hot slug has reached the far end of the heat exchanger. During all this time the hot slug was exposed only to zero-reference temperature fluid on the shell side. The temperature of the hot slug at time t_1 is therefore,

$$T_2' = T_2 e^{-t_1/\tau_t} \tag{28}$$

And the temperature of the fluid on the shell side is now Equation (29).

$$T_1' = \frac{\frac{T_2'}{\tau_t} C_0}{2V_0C_s} \tag{29}$$

After this time the hot slug will reverse the direction of its travel and will stay with the cold shell-side slug which has just entered. The two slugs, assumed to be of equal longitudinal thickness, will now travel to the left in Fig. 6 until they reach the outlet of the heat exchanger at the time,

$$t_2 = \frac{2L}{V_0} = \tau_{dt} \tag{30}$$

Heat transfer between the two slugs as they travel side by side will occur at a rate governed by the time constants of the two slugs and the decreasing temperature difference between them. Solution of the two differential equations, analytically or graphically, readily yields the temperatures of both slugs. We are interested only in the temperature of the shell-side slug. Let it be T_1'' at time t_2, the moment when the hot slug in the tube leaves the heat exchanger.

From C_0 we can calculate the pulse duration Δt. It is

$$\Delta t = \frac{4C_0}{\pi d^2 V_0 \rho c} \tag{31}$$

We now know the impulse response T_1, Fig. 7. The exponential part of the transient is described by

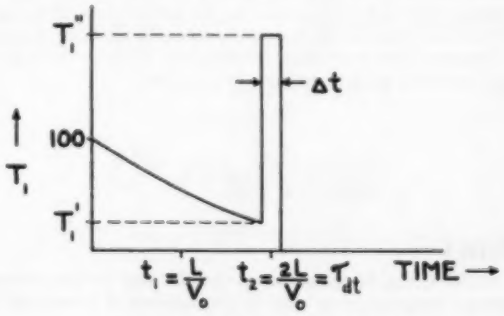


Fig. 7 Impulse response of Model 3

$$T_1 = 100\, e^{-t/2\tau_t}, \qquad 0 < t < t_2 \tag{32}$$

This is caused by spreading the dynamic transfer of energy from the hot slug with time constant τ_t over shell-side fluid twice the length of the heat exchanger.

To get the frequency response of this heater, relating T_1 to T_2, we take the Fourier transform of the impulse response. The result is

$$T_1(j\omega) = \frac{100\,(2\tau_t)}{1 + j2\omega\tau_t}\left[1 - e^{-\left(\frac{1}{2\tau_t} + j\omega\right)\frac{2L}{V_0}}\right] + \frac{T_1{}^{\sigma}}{j\omega}\, e^{-j\omega\frac{2L}{V_0}}\,[1 - e^{-j\omega\Delta t}] \tag{33}$$

Now,

$$e^{-j\omega\Delta t} = \cos\omega\Delta t - j\sin\omega\Delta t \simeq 1 - j\omega\Delta t \tag{34}$$

Substituting Equation (34) for the last exponential in Equation (33), and using $\tau_{dt} = 2L/V_0$, we get

$$T_1(j\omega) = \frac{100(2\tau_t)}{1 + j2\omega\tau_t}\left[1 - e^{-\left(\frac{1}{2\tau_t} + j\omega\right)\tau_{dt}}\right] + T_1{}^{\sigma}\Delta t e^{-j\omega\tau_{dt}} \tag{35}$$

Our assumption following Equation (23) allows us practically to ignore the quantity

$$e^{-\tau_{dt}/2\tau_t}\, e^{-j\omega\tau_{dt}}$$

in comparison to unity.

The frequency response of the heater is obtained from the unit-impulse response. We must therefore divide Equation (35) by $T_2 = K\Delta t$, the magnitude of the applied pulse. In addition we normalize to unity amplitude ratio at zero frequency. The frequency response becomes then

$$\frac{T_1(j\omega)}{T_2(j\omega)} = \frac{1}{1 + j2\omega\tau_t} + ke^{-j\omega\tau_{dt}} \tag{36}$$

with k a small quantity. Thus, on a logarithmic plot, at low frequencies, the first term of Equation (36) dominates, giving us the appearance of a first-order time-constant system. At higher frequencies, with the first term of Equation (36) smaller, the second term makes its appearance upon the logarithmic plot as a periodic undulation with peaks at

$$\omega = n(2\pi/\tau_{dt}),\ n = 0, 1, 2, \ldots \tag{37}$$

Masubuchi [8] does show such peaks for this type of heat exchanger. Note that the undulating nature of the frequency-response equation is produced by the second term of Equation (36), which arises in the parallel-flow part of the heat exchanger.

Equation (36), normalized for frequency on twice the tube-slug time constant, using Equation (38), becomes

$$\Omega_t = 2\tau_t\omega \tag{38}$$

$$\frac{T_1(j\omega)}{T_2(j\omega)} = \frac{1}{1 + j\Omega_t} + ke^{-j\frac{\tau_{dt}}{2\tau_t}\Omega_t} \tag{39}$$

Model 4

Model 4, Fig. 8, demonstrates the dilemma we face when the average temperature is used in the analysis of heat-exchanger dynamics. This model is found frequently in the nuclear-reactor-control literature [5, 6].

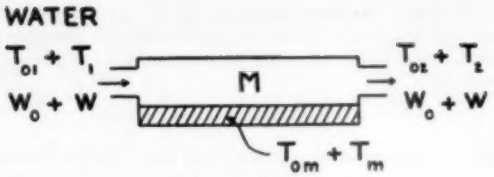


Fig. 8 Model 4

Heat is transferred from the metal plate with area A and uniform temperature $T_{om} + T_m$ to the water through film coefficient h. A linear temperature gradient is assumed to exist in the water between inlet and outlet.

The analysis, though illogical, usually proceeds as follows: The bulk temperature of the water in the heat exchanger is assumed to be equal to the arithmetic mean of the inlet and exit fluid temperatures. We therefore write the heat balance for small temperature perturbations about previous steady conditions,

$$hA\left[T_m - \left(\frac{T_2 + T_1}{2}\right)\right] - W_0c(T_2 - T_1) = Mc\frac{d}{dt}\left(\frac{T_2 + T_1}{2}\right) \tag{40}$$

We wish to find the frequency response between T_1 and T_2. For this we keep metal-plate temperature perturbation T_m at zero and substitute $j\omega$ for the derivative in Equation (40). Solving the resultant equation for the ratio T_2/T_1 we obtain the frequency response,

$$\frac{T_2(j\omega)}{T_1(j\omega)} = \left(\frac{2W_0c - hA}{2W_0c + hA}\right)\left(\frac{1 - j\omega\dfrac{Mc}{2W_0c - hA}}{1 + j\omega\dfrac{Mc}{2W_0c + hA}}\right) \tag{41}$$

From Equation (41) we find the amplitude ratio and phase angle to be

$$\text{AR} = \left[\frac{(2W_0c - hA)^2 + (\omega Mc)^2}{(2W_0c + hA)^2 + (\omega Mc)^2}\right]^{1/2} \tag{42}$$

$$\text{Phase angle} = -\arctan\frac{\omega Mc}{2W_0c + hA} - \arctan\frac{\omega Mc}{2W_0c - hA} \tag{43}$$

Equations (42) and (43) lead us to conclude that a very high-frequency temperature fluctuation at the inlet to the heat exchanger is transmitted to the outlet without attenuation but with 180-deg phase shift. Thus a sudden increase of inlet temperature T_1 will produce an equally sudden decrease of outlet temperature T_2 equal in magnitude to T_1. This is hardly a satisfactory conclusion.

We try next a new, but more logical analysis. The assumption that the bulk temperature of the water in the heat exchanger is equal to the arithmetic mean of the inlet and outlet fluid temperatures prescribes a model with the bulk temperature rigidly fixed to inlet and outlet temperature perturbations. A sudden increase of inlet temperature amounting to T_1 degrees will therefore without the slightest delay, in fact, immediately, increase the bulk temperature by $T_1/2$ degrees. While we may say that this is impossible physically, our major assumption has created

a mathematical model which does behave in this manner. But perhaps we will find the end result, the relationship between T_1 and T_2, to be more agreeable to us than that given by Equations (42) and (43).

Our heat-balance equation should be written

$$W_{gc}T_1 + (Mc)\left(\tfrac{1}{2}\frac{dT_1}{dt}\right) + hA\left[T_m - \left(\frac{T_2 + T_1}{2}\right)\right] - W_{gc}T_2 = (Mc)\left(\tfrac{1}{2}\frac{dT_2}{dt}\right) \quad (44)$$

The term

$$\tfrac{1}{2}(dT_1/dt)$$

shows that the changing inlet temperature, in accordance with our basic assumption, must be reflected in a changing bulk temperature at a rate half that at which T_1 is changing. And this rate of change of bulk temperature inescapably implies an additional rate of heat flow,

$$(Mc)[\tfrac{1}{2}(dT_1/dt)]$$

into the heat exchanger which must be accounted for. Finally, the incoming and outgoing heat flows must be balanced against the heat rate to bulk storage as reflected by the rate of change of outlet temperature; namely,

$$(Mc)[\tfrac{1}{2}(dT_2/dt)]$$

Equation (44) can be rewritten as

$$hA\left[T_m - \left(\frac{T_2 + T_1}{2}\right)\right] - W_{gc}(T_2 - T_1) = Mc\frac{d}{dt}\left(\frac{T_2 - T_1}{2}\right) \quad (45)$$

Equation (45) shows, logically so, that the net heat rate to bulk storage is proportional to the rate of change of the temperature difference between outlet and inlet temperatures. The net heat rate to bulk storage is not proportional to the sum of the rates of change of inlet and outlet temperatures, as implied, illogically, by Equation (40).

The frequency response relating T_2 to T_1, when derived from Equation (45), is given by the following two equations:

$$\frac{T_2(j\omega)}{T_1(j\omega)} = \left(\frac{2W_{gc} - hA}{2W_{gc} + hA}\right)\left(\frac{1 + j\omega\dfrac{Mc}{2W_{gc} - hA}}{1 + j\omega\dfrac{Mc}{2W_{gc} + hA}}\right) \quad (46)$$

$$\text{Phase angle} = -\arctan\frac{\omega Mc}{2W_{gc} + hA} + \arctan\frac{\omega Mc}{2W_{gc} - hA} \quad (47)$$

The amplitude ratio is unchanged from Equation (42).

Both Equations (40) and (45) give the same correct relationship between metal-temperature perturbations T_m and response of the outlet temperature T_2. But the frequency response between T_1 and T_2 is very different for the two equations. The polar diagram in Fig. 9 gives a qualitative illustration of this difference.

Clearly then, in considering stability of control for nuclear reactors, one should be aware of the limitations and possibly dangerous assumptions connected with the use of this model in a careless manner.

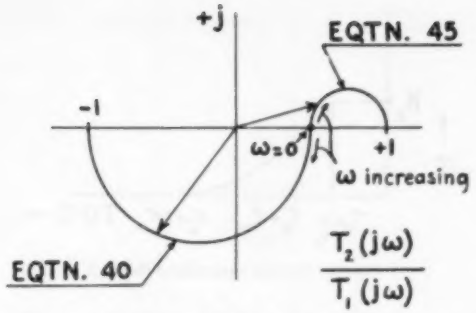


Fig. 9 Polar diagram for frequency response of Model 4

It is interesting to note the time constant τ_4, relating T_m to T_2:

$$\tau_4 = \frac{Mc}{2W_{gc} + hA} \quad (48)$$

This is appreciably less than the time constant τ_1, Equation (6), associated with the mixed fluid of Model 1. Mozley [9], in comparing Models 1 and 4, but as applied to two liquids flowing countercurrent, found Model 4 to be in better agreement with experimental data because of the smaller time constant.

Model 5

In our first four models we have ignored the heat-capacitance effect of metal walls in contact with the flowing fluids. In this, our last model, Fig. 10, we intend to consider the dynamic temperature interaction between a flowing fluid and its metal container.

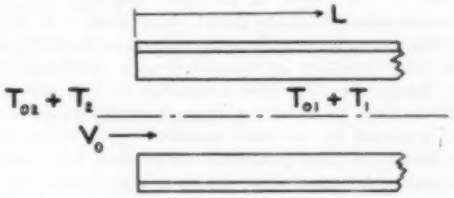


Fig. 10 Model 5. Section through insulated metal pipe.

We assume the outside of the pipe to be insulated. The pipe itself has finite thermal capacitance and negligible thermal resistance at right angles to its axis, but infinite resistance in the direction of the axis. This system was analyzed by Profos [3]. A brief explanation of the method of calculation used by Profos may be found in the book by Young [10]. A solution is listed in reference [10], giving temperature perturbation T_1 of a fluid at a downstream distance L when the inlet temperature is subjected to a sinusoidal oscillation of magnitude T_2 at circular frequency ω. Equation (49) is that solution

$$T_1 = T_2 e^{-a_1 L}\sin\omega\left(t - a_2\frac{L}{V_0}\right) \quad (49)$$

$$a_1 = \left(\frac{K_1}{V_0}\frac{\omega^2}{K_m^2 + \omega^2}\right) \quad (50)$$

$$a_2 = \left(\frac{K_1 K_m}{K_m^2 + \omega^2} + 1\right) \quad (51)$$

$$K_1 = \frac{4h}{d_1\rho_1 c_1} = \frac{1}{\tau_1} \quad (52)$$

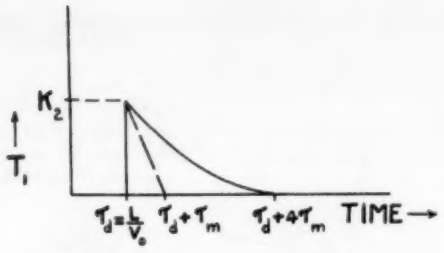


Fig. 11 Impulse response of Model 5

$$K_m = \frac{4d_1h}{(d_2^2 - d_1^2)\rho_m c_m} = \frac{1}{\tau_m} \tag{53}$$

Here τ_t is the time constant of the fluid and τ_m is the time constant of the metal wall. V_0 is fluid velocity.

While Equation (49) is not difficult to solve, it is time-consuming to plot various frequency-response curves for dynamic studies of systems which include pipes of assorted lengths and diameters conveying fluids at various speeds. Is it possible to construct approximate frequency-response curves quickly by utilizing easily calculated break-frequencies to locate the asymptotes for the logarithmic amplitude-ratio curves? Rizika [2] has published valuable data on temperature transients for gaseous fluids in pipelines. Can we convert his data to frequency-response curves? What major parameter should be selected on which to base our correlation?

Let us again approach this problem from the impulse-response point of view. Suppose we start a high-temperature slug of fluid down the tube. The impulse response T_1 will be approximately as shown in Fig. 11. After elapse of the dwell time τ_d, the temperature T_1 will show a pulselike response K_2, but of smaller magnitude than the initial pulse because of the exponential transfer of heat to the metal of the tube at a rate governed by the time constant τ_t associated with the fluid. Downstream temperature T_1, after responding to the pulse at time τ_d, will continue to register a higher temperature for a period largely governed by the time constant τ_m of the metal wall which determines the exponential rate at which heat is returned to the fluid. An exponential decays to 2 per cent of its initial value in a period of time equal to four time constants. We conclude therefore that duration of the transient response is from τ_d until $\tau_d + 4\tau_m$. The Fourier transform of this impulse response is

$$T_1(j\omega) = \frac{K_2\tau_m}{1 + j\omega\tau_m} e^{-\frac{\tau_d}{\tau_m}} \left[e^{-j\omega\tau_d} - e^{-4} e^{-j\omega(\tau_d + 4\tau_m)} \right] \tag{54}$$

Because e^{-4} is only 0.02 it follows that the magnitude of the second term in brackets may be neglected in comparison to the first term. Equation (54) is therefore practically equivalent to

$$T_1(j\omega) = \frac{K_2\tau_m}{1 + j\omega\tau_m} e^{-\frac{\tau_d}{\tau_m}} e^{-j\omega\tau_d} \tag{55}$$

We normalize again to unity amplitude ratio at zero frequency, yielding the frequency response,

$$\frac{T_1(j\omega)}{T_2(j\omega)} = \frac{1}{1 + j\omega\tau_m} e^{-j\omega\tau_d} \tag{56}$$

Now we are confronted with the choice of using either metal time constant τ_m or dwell time τ_d as our basic parameter. We suspect that knowledge of their relative magnitudes may be helpful in making this choice. Profos [3] gives us a good example. He compares two pipes, each 10 m long. Water flows in one of them, at a velocity of 0.5 m per sec. The other pipe carries low-pressure steam (1.1 atm abs, 200 C) at a velocity of 34 m per sec. The water pipe has an internal diameter of 19.6 mm and a 2.5-mm wall thickness. Corresponding dimensions for the steam pipe are 59 mm and 7.5 mm, respectively. The film coefficients of heat transfer are: For the water system, 2000 k-cal/m²-deg C-hr; for the steam system, 150 k cal/m²-deg C-hr. Specific heat and density of the steam are 0.5 k cal/kg-deg C and 0.5 kg/m³, respectively. For the pipe metal these quantities are approximately 0.13 k cal/kg-deg C and 7800 kg/m³. Calculated time constants are shown in Table 1.

Table 1 Time constants for steam and water pipes

	Water	Steam
Dwell time, τ_d, sec	20	0.29
Metal, τ_m, sec	4.8	214
Fluid, τ_t, sec	0.88	8.85

Table 1 immediately leads us to test τ_d for the water pipe and τ_m for the steam pipe as the characteristic parameters. We shall therefore plot the frequency-response data reported by Profos for the water pipe in polar co-ordinate form on our rectangular co-ordinate system but normalized to unity amplitude ratio at zero frequency. For the normalized circular frequency we will be consistent with Model 2. The water pipe, then, will be normalized on the circular frequency defined by

$$\Omega_d = \frac{\tau_d}{2}\omega \tag{57}$$

The filled-in circles in Fig. 12 are the result. At $\Omega = 1$, the characteristic frequency, we find $\omega = 2/\tau_d = 0.1$ rad/sec. If the water pipe had been a lumped first-order system, as implied by the amplitude of Equation (56), we would have seen the characteristic break frequency at $\omega = 1/\tau_m \simeq 0.2$ rad/sec. But the water pipe is not a lumped first-order system. That is also shown by the increase with frequency of the slope of the amplitude-ratio curve. First, the curve matches the first-order dashed curve in Fig. 12 marked τ_1. Then, between $\Omega = 1.5$ and 3 it fits the second-order curve marked $\tau_1 + \tau_2$.

The phase angle, to our considerable satisfaction, agrees perfectly over the range shown in Fig. 12 with the curve of a first-order system plus dead time as described by Equation (56). The characteristic frequencies involved are $\omega = 2/\tau_d = 0.1$ rad/sec, as before, plus the one caused by dead time, or dwell time; namely, $\omega = 1/\tau_d = 0.05$ rad/sec. Corresponding normalized frequencies, defined by Equation (57), are $\Omega = 1$ and $\Omega = 0.5$.

For the gaslike fluid, steam, we suspect that the small value of τ_d in comparison with τ_m will make its dynamic effect negligible. But data provided by Rizika [2] show that a single time constant will not suffice for any but the shortest pipe lengths when gaseous fluids are involved. However, for short pipe lengths τ_m is indeed adequate.

By plotting the transient remainders of the family of transient-response curves, published by Rizika, on semilog paper it is possible to extract without much difficulty the time constants which may be approximately associated with the transients. An earlier paper by the author [17] has shown that two or three time constants plus dead time can be combined in various proportions to yield an essentially identical step transient. The only requirement, within limits, appears to be that the summation of all time constants plus dead time, if any is used, be a constant for a particular step transient. With this thought in mind the combination of time constants extracted from the Rizika transients were plotted as shown in Fig. 13. The dashed lines provide a convenient approximate relationship between time constants and pipe length. The time-constant scale in Fig. 13 is nor-

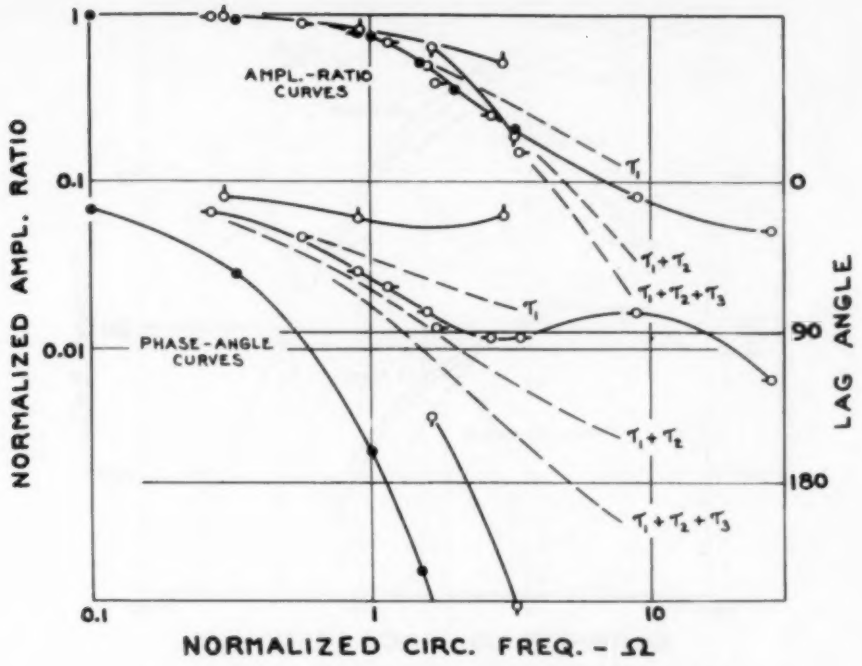


Fig. 12 Frequency response of metal pipes (based on Profos' data)

Steam Pipe (steel) 59 mm ID × 7.5-mm wall.
(1) L-P steam: 1.1 ata, 200 C, 34 m/sec, τ_m = 214 sec ◑, 2.5-m long pipe, L/D = 0.835, τ_1 = 303 sec, Ω = (303 sec) (ω rad/sec)
○, 10-m long pipe, L/D = 3.33, τ_1 = 570 sec, Ω = (570 sec) (ω rad/sec)
♀, 40-m long pipe, L/D = 13.3, τ_1 = 1640 sec, Ω = (1640 sec) (ω rad/sec)
(2) H-P steam: 100 ata, 400 C, 34 m/sec, τ_m = 10.7 sec ○, 50-m long pipe, L/D = 3.0, τ_1 = 26.8 sec, Ω = (26.8 sec) (ω rad/sec)
Water Pipe (steel): 19.6 mm ID × 2.5-mm wall
Water: 0.5 m/sec, τ_m = 4.8 sec
●, 10 m-long pipe, τ_d = 20 sec, τ_0 = 20 sec, $\Omega = (\tau_d/2)\omega$ = (10 sec)-(ω rad/sec)

malized on τ_m, while the abscissa is a dimensionless pipe distance normalized on the quantity D, in feet, also used by Rizika, and defined by

$$D = \frac{\text{heat-capacitance flow rate}}{\text{conductance for heat flow between fluid and pipe wall, per unit length of pipe}} = \frac{W_0 c_t}{\pi d h} \tag{58}$$

Finally, we choose as our major parameter for the steam pipe the major time constant from Fig. 13; namely, τ_1, which is a function of τ_m defined by

$$\tau_1 = \left(1 + \tfrac{1}{2}\frac{L}{D}\right)\tau_m \tag{59}$$

The Profos curves for steam pipes, which include the system under discussion, are shown replotted in Fig. 12 after normalizing to τ_1. With Rizika data extending only to a normalized pipe length of 5 it becomes necessary to extrapolate these data far out to a normalized pipe length of 13 to enable the high-pressure (100 atm abs, 400 C) long (50 m) pipeline data of Profos to be placed in Fig. 12.

Figs. 12 and 13 show that more time constants are necessary to approximate the frequency response of the longer steam lines. These additional time constants were chosen, again approximately

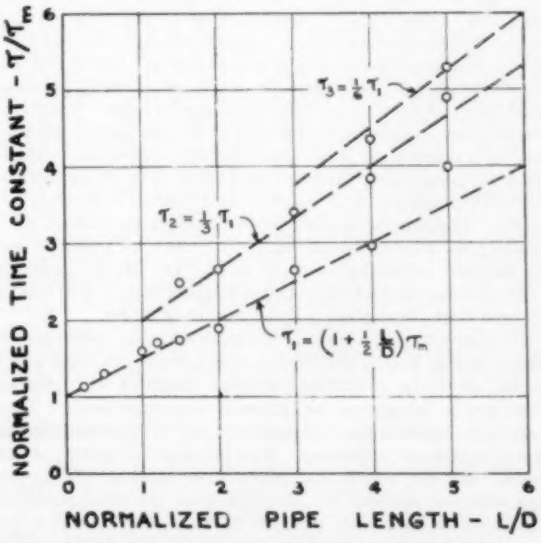


Fig. 13 Time constants for gaseous flow in metal pipes (based on Rizika data)

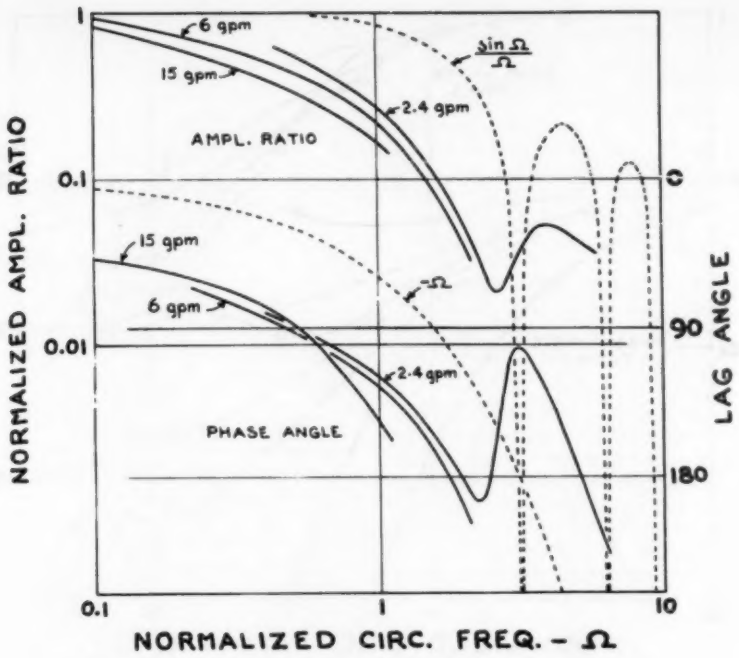


Fig. 14 Frequency response of a heat exchanger (based on DeBolt data)

and for simplicity, as simple fractions of the major time constant; namely,

$$\tau_2 = \tau_1/3 \tag{60}$$

$$\tau_3 = \tau_1/6 \tag{61}$$

Additional Observations

DeBolt [11] has tested a heat exchanger with an unusually large sinusoidal pneumatic-pressure fluctuation applied to the control valve admitting saturated steam to the shell. This pressure was varied between 10 and 14 psig, providing a 2-psi amplitude. The effective working range of the $^3/_4$-in. control valve was the standard 3 to 15 psig between full closed and full open. The heat-transfer area of the heater was 5.4 sq ft. Water flowed through the tubes in three passes, each about 4 ft long. Each pass had three $^5/_8$-in., 16 BWG admiralty tubes. Saturated steam to the control valve was held at 65 psia and 97 per cent quality. Condensate level was carefully controlled. Outlet-water temperature fluctuations were picked up by a bare thermopile with negligible time constant and put on a Sanborn recorder. Frequency-response curves obtained by DeBolt at 6 and 15 gpm, with corresponding dwell times τ_d of 4 and 1.6 sec, are redrawn, normalized to $\tau_d/2$, in Fig. 14. He obtained the curve showing undulation at higher frequencies at a flow rate of 2.4 gpm, with corresponding dwell time τ_d of 10 sec.

Frequency-response tests on the industrial heat exchanger reported in Fig. 5 with the filled-in circles were performed under actual operating conditions, without upsetting the process. Data had to be obtained at extremely low signal levels. They were recovered from the background "noise" by machine-correlation-computation techniques. This accounts for scatter of the points. An appreciable and unavoidable unknown thermowell time constant explains the deviation from the other points in Fig. 5. The data were included because they represent the largest dwell time τ_d available to the author; namely, 30 sec, ten times the smallest shown. This heat exchanger also differed from the others reported in that hot liquid inside the tubes was used to heat an organic liquid in the shell.

The Lees and Hougen data shown in Fig. 5, together with those discussed in the preceding paragraph, relate tube-exit temperature fluctuations to tube-flow perturbations. This information confirms our thesis, proposed under Model 2, that dwell time τ_d may be used as the major characteristic parameter for describing these heat exchangers; that is, for relating either tube-flow variations or shell-temperature perturbations to tube-outlet-temperature fluctuations.

It would appear that the class of heat exchangers shown in Fig. 5 can be described, at least to a first approximation, by means of the frequency-response curves in Fig. 15. These curves contain a major time constant τ_1, plus a secondary time constant τ_2, and a dead time τ_0, all based on the dwell time τ_d, as defined by the following equations:

$$\tau_1 = T_d \tag{62}$$

$$\tau_2 = \tau_d/4 \tag{63}$$

$$\tau_0 = \tau_d/4 \tag{64}$$

For example, Catheron, Goodhue, and Hansen [12] report equivalent time constants for their heat exchanger and we compare these in Table 2 with values derived from Equations (62) to (64). (NOTE: All time constants in Tables 2, 3, and 4 are given in seconds.)

Table 2 Catheron, Goodhue, Hansen data compared

At 20 gpm	τ_1	τ_2	τ_0	$\Sigma\tau$
Shell steam-flow perturbation	12	4	3	19
Shell steam-pressure perturbation	8	2	3	13
From Equations (62) to (64), with τ_d = 11 sec	11	3	3	17

Another example is provided by the Lees and Hougen [13] data with tube-liquid flow changes in a heat exchanger quite similar to the one reported in Table 2. Their figures are compared in Table 3.

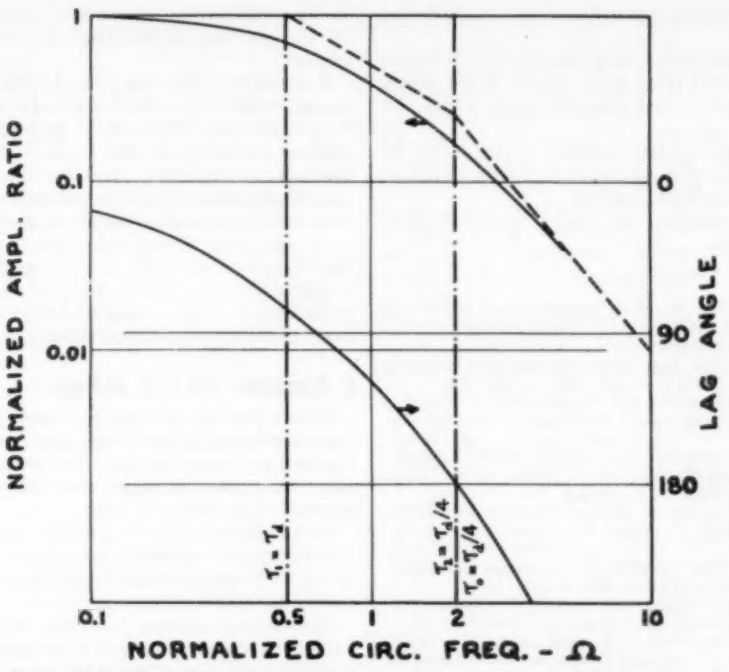


Fig. 15 Approximate frequency response of heat exchangers (based upon data in Fig. 5)

Table 3 Lees and Hougen data compared

At 30 gpm	τ_1	τ_2	τ_0	$\Sigma\tau$
Tube-liquid flow changes	7.6	1.41	1.75	10.8
From Equations (62) to (64), with τ_d = 7.2 sec	7.2	1.8	1.8	10.8

Cohen and Johnson [14] used an 11.6-ft-long single-pass concentric-tube heater, made of 1-in. and 2-in. brass pipe, with steam on the shell side. Their figures are compared in Table 4.

Table 4 Cohen and Johnson data compared

Water at 1.44 lb/sec	τ_1	τ_2	τ_0	$\Sigma\tau$
Shell steam-flow perturbation (temp obtained from sat stm press)	3	1	1	5
From Equations (62) to (64) with τ_d = 3 sec	3	0.75	0.75	4.5

An example of transient-response correlation is provided by Rizika in a second paper [15] in which he computes the transient response of a Ross heater No. 812 (8 in. in diameter and 12 ft long, consisting of 72 $^5/_8$-in-OD, No. 18 BWG brass tubes). This heater has been designed to heat, in the steady state, 6300 gal of water per hr by condensing 5200 lb of steam (at 10 psig) per hr on the outer surface of the heat-transfer tubes. The dwell time under the specified operating conditions is listed by Rizika as τ_d = 5.65 sec. Hence, we would expect our major time constant to be $\tau_1 = \tau_d$ = 5.65 sec, with a secondary time constant and dead time each equal to 1.41 sec. Rizika calculates his transient under the following extreme conditions: Water is allowed to pass through the tubes at rated flow until the metal of the heater is at equilibrium with the cold water. All this time no steam is admitted. Suddenly the steam is allowed to flow. Rizika's calculated transient, which shows no dead time and about 0.5-sec secondary time constant may be approximated with a maximum-slope line at the point of inflection. This line cuts the new steady-state condition line at almost 5 sec and the initial-condition line at 0.5 sec. Rizika's calculated transient-response curve follows the maximum-slope line for 2 sec. To a first approximation we represent the Rizika transient by a 4.5-sec major time constant plus 0.5-sec secondary time constant, producing $\Sigma\tau$ = 5 sec. The $\Sigma\tau$ based on τ_d and Equations (62) to (64) would have given us 5.65 sec plus twice 1.41 sec or 8.5 sec. Apparently even such severe test results can be readily predicted with reasonable accuracy by using the data and methods provided in this paper.

We add an observation concerning the interaction of hot gases with metal walls by quoting Cima and London [16]: "Another saving feature from a control point of view is that . . . the response to a flow step change is essentially complete within a time interval $t = \tau_d$, and this is substantially faster than the response to a temperature step change . . . Thus, as step-function changes in flow will also be accompanied by step-function changes in inlet temperature, one can expect the 50 or more times slower response to the inlet temperature changes to control the over-all behavior."

For our final observation we cite the point made by Rizika [2] in the last paragraph of his paper to the effect that it is entirely possible for a heat-exchanger and long-pipe combination to act dynamically as if the heat exchanger were not even present. We believe that this consideration may well be of importance to nuclear-reactor-control engineers especially in those cases where shielding necessitates long pipelines between the reactor core and the next major point of heat transfer. The data presented in this paper should provide a simple and ready means for determining when such a difficult control condition is likely to arise.

Conclusions

1 Basic dynamic characteristics of heat exchangers can be found by analysis of very simple mathematical models.

2 A single basic parameter may be used to describe dynamic characteristics of entire families of heat exchangers.

3 Parameters which may be so used are dwell time of the

fluid within the heat exchanger and the metal and fluid time constants.

4 A single dynamic parameter easily adjusted to fit numerous operating conditions should prove to be helpful in the design, purchase, use, classification, and standardization of heat exchangers.

5 Appreciable amounts of data produced by apparently diverse researches on heat-exchanger dynamics may be correlated and presented in more condensed fashion.

6 The average temperature in heat-exchanger analysis should be used with caution.

References

1 J. R. Newman, "The World of Mathematics," Simon and Schuster, New York, N. Y., 1956, vol. 4, p. 2160: Morse and Kimball, "How to Hunt a Submarine."

2 J. W. Rizika, "Thermal Lags in Flowing Systems Containing Heat Capacitors," TRANS. ASME, vol. 76, 1954, pp. 411–420.

3 P. Profos, "Die Behandlung von Regelproblemen vermittels des Frequenzganges des Regelkreises," PhD thesis, Technical University, Zurich, 1943.

4 Y. Takahashi, "Transfer Function Analysis of Heat Exchange Processes," "Automatic and Manual Control," edited by A. Tustin, Academic Press Inc., New York, N. Y., 1952, p. 235.

5 M. A. Schultz, "Control of Nuclear Reactors and Power Plants," McGraw-Hill Book Company, Inc., New York, N. Y., 1955, p. 138.

6 John MacPhee, "The Stability of Multipass Reactors," *Nuclear Science and Engineering*, vol. 4, 1958, p. 203.

7 E. M. Grabbe, Simon Ramo, and Dean E. Wooldridge, editors, "Handbook of Automation, Computation and Control," vol. 1, John Wiley & Sons, Inc., New York, N. Y., 1958, pp. 22–52 to 22–60.

8 Masami Masubuchi, "Dynamic Response and Control of Multipass Heat Exchangers," TRANS. ASME, series D, JOURNAL OF BASIC ENGINEERING, vol. 82, 1960, pp. 51–65.

9 J. M. Mozley, "Predicting Dynamics of Concentric Pipe Heat Exchangers," *Industrial and Engineering Chemistry*, vol. 48, 1956, pp. 1035–1041.

10 A. J. Young, "An Introduction to Process Control System Design," Longmans, Green and Company, New York, N. Y., 1955, pp. 268–277.

11 R. R. DeBolt, "Dynamic Characteristics of a Steam-Water Heat Exchanger," MS thesis, University of California, Berkeley, Calif., 1954.

12 A. R. Catheron, S. H. Goodhue, and P. D. Hansen, "Control of Shell and Tube Heat Exchangers," ASME Paper No. 59—IRD-14.

13 Sidney Lees and J. O. Hougen, "Pulse Testing a Model Heat Exchange Process," *Industrial and Engineering Chemistry*, vol. 48, 1956, pp. 1064–1068.

14 W. C. Cohen and E. F. Johnson, "Dynamic Characteristics of Double-Pipe Heat Exchangers," *Industrial and Engineering Chemistry*, vol. 48, 1956, pp. 1031–1034.

15 J. W. Rizika, "Thermal Lags in Flowing Incompressible Fluid Systems Containing Heat Capacitors," TRANS. ASME, vol. 78, 1956, p. 1412.

16 R. M. Cima and A. L. London, "The Transient Response of a Two-Fluid Counterflow Heat Exchanger—The Gas-Turbine Regenerator," TRANS. ASME, vol. 80, 1958, p. 1173.

17 H. Thal-Larsen, "Frequency Response From Experimental Nonoscillatory Transient-Response Data," *Trans. AIEE*, Part II, "Applications and Industry," 1955, pp. 109–113.

DISCUSSION

Donald J. Bergman[3]

One of the dynamic lags in a heat exchanger results from the necessity of measuring outlet temperature (for control) at a point fixed by the mechanical design of the exchanger. Ordinarily, this is close to the outlet flange. It is obvious that taking the measurement at a point 10 ft down the pipe from the exchanger outlet with the heated fluid moving at a velocity of 5 ft/sec adds 2 seconds to this particular time lag.

In the firing control of oil heaters it was found out back in the 1930's, mostly by intuition and trial, that inserting the control thermocouple perhaps 10 or 12 tubes ahead of the outlet shortened the apparent time lag and improved heater outlet temperature control.

It would seem that this particular time lag of a heater or exchanger could be approximated by an estimate of the tube length back to the center of gravity of the area between the two temperature curves and the time required for the fluid to flow from this point to the outlet. Location of a control couple at this point would probably result in unsatisfactory variations of outlet temperature with load, consequently, the optimum location would be at some point between this center of gravity and the outlet. This leads to the conclusion that, where multiple exchangers or multipass units are used, improved control might be obtained by locating a couple between the third and fourth shell, or inserted in the channel ahead of the last pass of tubes.

[3] Chief Engineer, Engineering and Development Department, Universal Oil Products Company, Des Plaines, Ill. Mem. ASME.

C. Concordia[4] and F. P. deMello[5]

Correct simulation of transient flow and heat-transfer phenomena is becoming increasingly important as automatic control is applied to processes in which these phenomena play a dominant role. The author's treatment of the subject is an important contribution in this field.

We have noted in particular the discussion of Model No. 4 and the author's suggestion for use of a better or more logical model. We cannot entirely agree with his treatment of this case for the following reasons:

Mathematical relations describing the performance of Model 4 are partial differential equations in space and time. A feasible solution of such partial differential equations can be obtained by eliminating the space variable through lumped-constant representation whereby the space is broken up into finite nodes. A true representation valid for all frequencies would require an infinite number of nodes. However, in most actual problems the frequency band of interest is limited, and a finite-number-of-node representation is adequate.

The partial differential equation describing the system of Model 4 can be written as

$$Mc\,\frac{\partial T}{\partial t} + cW\,\frac{\partial T}{\partial x} - hA(T_m - T) = 0 \qquad (65)$$

where

M = mass per unit length
c = specific heat of fluid
W = mass flow per unit of time
T = temperature of fluid (function of x and t)
T_m = temperature of metal (function of x and t)
hA = heat-transfer coefficient per unit of length

Equation (65) can be approximated by an ordinary differential equation relating variables defined at discrete points, x_1 and x_2, with corresponding temperatures of T_1 and T_2. When these points are close enough together the differential equation applies to an elemental volume for which the average temperature can be approximated as $\frac{T_1 + T_2}{2}$. This procedure results in equation (40) of the paper, which is commonly used.

The fallacy is to use only one node to simulate the heat exchanger. Actually with 3 nodes the transfer function would be

$$\frac{T_2(p)}{T_1(p)} = \left(\frac{2W_0c - hA/3}{2W_0c + hA/3}\right)^3 \frac{\left(1 - \dfrac{Mc/3}{2W_0c - hA/3}\,p\right)^3}{\left(1 + \dfrac{Mc/3}{2W_0c + hA/3}\,p\right)^3} \qquad (66)$$

[4] Electric Utility Engineering Section, General Electric Company, Schenectady, N. Y. Fellow ASME.

[5] General Electric Company, Schenectady, N. Y.

and with n nodes the transfer function would be:

$$\frac{T_2(p)}{T_1(p)} = \left(\frac{2W_{0}c - \frac{hA}{n}}{2W_{0}c + \frac{hA}{n}}\right)^n \frac{\left(1 - \frac{Mc/n}{2W_{0}c - hA/n}p\right)^n}{\left(1 + \frac{Mc/n}{2W_{0}c + hA/n}p\right)^n} \quad (67)$$

The form of equation (67) shows that a pseudo transport delay is involved since the approximation to a transport lag is

$$e^{-pT} = \frac{\left(1 - p\frac{T}{n}\right)^n}{\left(1 + p\frac{T}{n}\right)^n} \quad (68)$$

When $hA/n = 0$, i.e., when there is no heat transfer from metal to fluid, expression (67) is exactly of the form of expression (68), moreover this is the actual case since, with no heat transfer a change in T_1 propagates to T_2 as a pure transport lag.

Typical of the finite-number-of-node approximation of a partial differential equation in space and time is the so-called ringing effect, where the output variable oscillates at a frequency proportional to the number of nodes. Where this frequency is high compared to the frequency of other phenomena of interest it does not introduce objectionable errors since it can be recognized for what it is.

We were also interested to note that the author dismisses the analysis of the steady state as relatively simple. Unfortunately this is not always the case. Indeed in many cases an adequate analysis of the steady-state performance may require much more effort than that required for adding the dynamic features. This is because in the design of control systems the required accuracy of representation of steady-state phenomena may be considerably greater than that of dynamic effects, and it is often necessary to consider nonlinear effects. Once the actual modes of operation are properly understood, the dynamic effects can sometimes be approximated. Even when improvements in performance during transients is the primary object of a study, the steady-state representation may still have major importance. An example occurs in the design of so-called noninteracting controls.

Arvid Hempel[6]

The author's idea of trying to characterize the dynamic behavior of heat exchangers by a single parameter is a very attractive one and is well worth investigating. In my opinion, however, it seems strange that a liquid-liquid heat exchanger can be described by the same transfer function as a steam-liquid heat exchanger. In this paper only experimental results from one liquid-liquid heat exchanger are presented, and these are those which do not fit in very well with the others. It is not clearly stated on which parameter these data are normalized; whether it is half the dwelltime of the shell flow or the tube flow. According to the treatment of model 3, one should expect that the time constant of the tube fluid would be used.

I would suggest that the two systems mentioned above should be treated separately.

Concentrating on the steam-liquid system: From a theoretical point of view nothing can be said about any time constants in the system from knowledge of the dwelltime of the liquid. For instance, doubling of the length of the heat exchanger does not affect any time constants. However, sound static design of heat exchangers requires that the ratio of the dwell time τ_d and the time constant of the liquid τ_l be fairly constant for all heat exchangers of this type. Within limits it might be possible that the time constant associated with the tubes, is related to the liquid time constant and therefore also to the dwell time. This may be the reason that a treatment based on the dwell time perhaps will be adequate in determining the dynamic behavior of industrial heat exchangers. According to the foregoing reasoning, one should be careful in selecting experimental results for further checking of this theory and not use data from heat exchangers of abnormal static design.

To illustrate this reservation, I have checked the author's result against frequency response data from an industrial tube and shell single pass heat exchanger, with eighteen Cr-Ni tubes in the shell.

Condensing steam was used to heat and boil a solution flowing at a rate of 40 dm³/min. The length of the heat exchanger was 7900 mm, and the dwell time 1.2 min. The result is shown together with the author's Fig. 15.

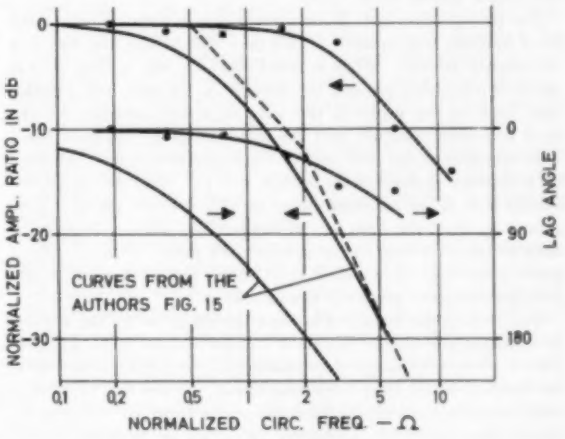


Fig. 16

It is seen that the agreement is very poor. This may be explained by the fact that, because boiling occurs in the heat exchanger, the ratio τ_d/τ_l has not the normal value for the heat exchanger.

The first part of the paper represents a new approach of deriving the transfer functions for heat exchangers. I cannot, however, clearly see the relation between this and the results obtained in the last part. None of the derived transfer functions are of the form found empirically to describe the process.

Obviously, the important task of this approach is to derive the impulse response of the systems. Even if only an approximate impulse response of the given model is wanted, it seems to me that the author has gone too far with the simplifications in deriving the impulse responses for some of his models. This is pointed out in the following remarks.

Model 2

First the author here has made an unnecessary assumption that the temperature distribution of the water is linear. As the system is linear, the impulse in the steam temperature may be considered to occur when the temperature is uniform throughout the system, thus the temperature distribution will not affect the result at all.

The response given in Fig. 4 cannot be correct because, when the impulse has disappeared, the water will lose heat at a rate governed by the time constant τ_w of the water, until $t = \tau_d$, then it drops to zero. Analytically, this may be expressed

$$T_2(t) = U(\tau_d - t)e^{-t/\tau_w}$$

[6] Chr. Michelsens Institute, Dept. of Applied Physics, Bergen, Norway.

where $U(\tau_d - t)$ is the unit step function.
Thus

$$U(\tau_d - t) = \begin{cases} 1 & \text{for } t < \tau_d \\ 0 & \text{for } t > \tau_d \end{cases}$$

And by Laplace transforming this, the transfer function

$$\frac{T_2}{T_1}(s) = \frac{1}{\tau_w s + 1}\left[1 - e^{-\left(s+\frac{1}{\tau_w}\right)\tau_d}\right] \text{ is obtained.}$$

The considerations made concerning the response to the velocity, however, seem to be correct.

Model 3

This model is not easy to analyze, and I cannot see that the author's result represents the response of the system with an acceptable degree of accuracy.

For his equation (24) to be correct, the system initially must be of uniform temperature. This may well be assumed but it is not clearly stated. What is overlooked by the author is that the shell fluid, after passing the impulse in the tube, will transfer heat back to the water in the tube at a rate governed by the shell flow time constant and the temperature of the shell fluid. This means that the first part of the response curve is described by a function depending on both τ_s and τ_t. This will affect the second part of the response curve as well, because the heat thus gained by the tube fluid in the counterflow part is transferred back to the shell fluid in the parallel flow part. This will give a gradual decrease of the shell fluid outlet temperature after the pulselike response given by the author.

This is the main objection to the response given by the author. In addition the author has used a pulse instead of an impulse. This is perhaps of minor importance but the effect on the transfer function could have been mentioned. I also find it strange that the author calculates the pulse width from the value of C_0, which then has to be estimated. It would be more natural to calculate C_0 from a given pulse width.

Model 5

The impulse response of this model, given in Fig. 11, is also a simplified one, because actually the heat gained by the fluid will depend on the temperature of the fluid as well as the time constant τ_m, and the fluid temperature will vary and make the calculation complicated.

However, accepting the exponentially decaying response given, it is not necessary to state the duration of the response. In fact, if the response is written

$$T_1 = U(t - \tau_d)e^{-t/\tau_m}$$

where

$U(t - \tau_d)$ is the unit step function

then Fourier transformation at once gives the equation (55) of the author.

A. L. London[7]

This discusser agrees with the author's view that a normalized time makes for considerable generalization in describing the response of heat exchangers. In another paper[8] this problem is considered from the dimensional analysis point of view. These results combined with some analog solutions plus some very special case analytical solutions lead to the conclusion that normalized time θ^* and normalized wall capacitance $\bar{C}_w^*$ in *combination* provide even greater generality. This was not evident in the author's treatment because he deals with a class of exchangers (thin metal wall) where the wall capacitance parameter $\bar{C}_w^*$ approaches zero. Although results are reported[8] in the form of transient-response curves, these parameters, as well as the others considered,[8] will apply also to the frequency response as noted by the author in reference [17].

Table 2[8] should prove to be useful in the classification of heat exchangers.

[7] Professor of Mechanical Engineering, Stanford University, Stanford, Calif. Mem. ASME.

[8] A. L. London, F. R. Biancardi, and J. W. Mitchell, "The Transient Response of Gas-Turbine-Plant Heat Exchangers—Regenerators, Intercoolers, Precoolers, and Ducting." TRANS. ASME, series A, *Journal of Engineering for Power*, vol. 81, 1959, pp. 433–448.

Y. Takahashi[9]

The author is to be congratulated for presenting a new, simple approach in heat exchanger dynamics fitted for control engineers' use.

There are many standpoints in dealing with the analysis and design of heat exchangers. Also there are so many varieties in types, size, constructional details, and operating modes.

However, the information required by control engineers on heat exchangers is rather simple. The transfer function, especially the dynamic part of the transfer function, is the information they need. The static part of the transfer function, the d-c gain, is important but may be found elsewhere, from steady-state design calculation, for example. Even when the system is nonlinear, which is almost always the case if one closely investigates a system, the approximate linear transfer function will be the information to be used, although cautiously, in dynamic system design.

As is pointed out by the author, one of the difficulties in deriving the transfer function or frequency response is in the complexity of the system. Quite often it is a distributed parameter system requiring partial differential equations to describe the system.

Here is the beauty of the author's simple, unified approach. It is a thinking experiment, not an actual test. Through the thinking experiment one can visualize an impulse response, as described in the paper for several models, which in turn may be expressed in a mathematical transfer function form. One must have intuition and experience in order to eliminate factors of secondary importance, leaving only major factors. A factor such as heat storage at one place may be neglected in one case, but not so in other cases.

Thus the accuracy of the result may be influenced by the assumptions one makes in his derivation process. Also, the required accuracy, which is dependent upon the problem at hand, must be considered. Thus the thinking experiment process itself may be regarded as a feedback process including the control engineer in the closed loop. Through that feedback process, use of the impulse-response principle may yield the frequency response with satisfactory accuracy on a band-width of importance for system design.

Author's Closure

The author thanks the discussers for their interesting and valuable contributions further elucidating the problem of heat-exchanger dynamics.

Dr. Bergman's pertinent observations show that for many years dwell time has been recognized as a governing parameter in the control of heat exchangers. It is surprising, therefore, that normalized time and normalized frequency, both based upon dwell time, were not used earlier to unify apparently diverse researches on the dynamics of heat exchangers.

[9] Professor of Mechanical Engineering, University of California, Berkeley, Calif.

Messrs. Concordia and deMello have shown a line of reasoning for arriving at the commonly used Equation (40) of the paper which starts out rather logically with a partial differential equation. In their development they at least point out some of the unsatisfactory features, such as the fallacy of using only a single node, described by Equation (40), and the occurrence of the so-called ringing effect when a finite number of nodes is used. While such an approach is preferable to a derivation of Equation [40) based upon the average temperature, the danger remains that Equation (40) will be used uncritically.

Let us remember that Model 4 is supposed to describe a nuclear-reactor fuel element. We wish to relate the inlet and outlet fluid temperatures. Note that with constant heat generation in the fuel element a small increase in inlet temperature will eventually be reflected in an equivalent rise of the outlet temperature, at least to a first approximation. This would not be the case in Model 2 with constant steam temperature. For an adequate description of Model 4 we must therefore have unity amplitude ratio at zero frequency for the function relating inlet and outlet temperatures. The function must also contain a dead time or dwell time, and a time constant. The step-transient response will then be more realistic, without the misleading "reverse-reaction effect" associated with Equation (40).[10] A model based upon Model 1 plus dead time should therefore provide a more satisfactory model for relating fluid temperatures, in better agreement with observed results, free from such effects as the reverse reaction, ringing, and steady-state gain less than unity.

Messrs. Concordia and deMello have misunderstood the author when they remark: "We have noted . . . the author's suggestion for use of a better or more logical model." More logical, yes; but better, definitely not! The author hopes that a careful rereading of his remarks in connection with the "new, but more logical analysis" will convince them that he has been careful not to pass judgment on the merits of the "more logical" model.

It is the author's impression that steady-state analysis is generally regarded as being much simpler than transient analysis. The greater part of engineering education, for example, is still devoted to steady-state analyses, largely because we know so little about the transient state. In this connection there are those who maintain that thermodynamics as now taught is misnamed, that it should be called thermostatics. It is not until students have acquired maturity and mathematical skill that they are able to cope with some aspects of the transient behavior of systems. The author is convinced that Messrs. Concordia and deMello know this, but that they want to warn the unwary reader that even the steady-state analysis can have its pitfalls. And such a pitfall is the steady-state gain less than unity for Equation (40) relating T_1 and T_2.

Mr. Hempel discusses Models 2, 3, and 5. Starting with Model 2, he remarks that it is not clearly stated on which parameter the data for the liquid-liquid heat exchanger in Fig. 5 are normalized. The author apologizes for this oversight. The reported dwell time of 30 seconds refers to the liquid inside the tubes. Boiling of the shell-side liquid apparently produced a sufficiently uniform shell-side temperature for Model 2 assumptions to be applicable. The author's comments under "Additional Observations" explain why the data for the liquid-liquid heat exchanger deviate from the other data in Fig. 5 and give the reason for nevertheless including them on Fig. 5.

[10] Of interest is a comment concerning initial negative values made by Professor H. M. Paynter on page 242 of "A Palimpsest on the Electronic Analog Art," edited by him and printed by George A. Philbrick Researches, Inc., Boston, Mass., 1955: ". . . the initial negative values . . . [are] characteristic of the method and not of the natural stream." Fig. 15(d) on page 241 of that publication shows the reverse reaction effect. And we note in passing that the reverse reaction actually does occur in boiler-level control, and also in airplane control, for example.

Mr. Hempel warns that one should ". . . not use data from heat exchangers of abnormal static design." If Mr. Hempel means that the heat-exchangers reported on under Model 2 are of "abnormal static design," then a very large number, perhaps the majority of heat exchangers, are indeed operated under "abnormal static" conditions. By not using data from such heat exchangers one would effectively ignore a large class of industrial heat exchangers.

The fact that Mr. Hempel's heat exchanger with shell-side steam boiling the liquid inside the tubes does not fit Model 2 data is not surprising because the assumptions of Model 2 do not apply. Nevertheless, it is interesting to note that dividing the normalized frequency of the boiling heat exchanger by about 5 will shift its amplitude-ratio curve into a beautiful fit with the amplitude-ratio curve from Fig. 15. And the discrepancy in phase shift at the high-frequency tip of the phase curve is then about 30 degrees, a small error in comparison to the basic similarity between the two sets of curves. A dwell time about $^1/_5$ the reported value of 1.2 minutes can account for such a shift. At this point, in the absence of sufficient information, the author is tempted to speculate as follows: Can the frequency shift be accounted for by an effective liquid length within the tubes equal to $^1/_5$ the tube length, with the remainder of the space occupied by vapor?

The author did not make the assumption that the temperature distribution of the water is linear as indeed such an assumption is unnecessary. His statement ". . . a steady average temperature gradient. . ." does not imply that.

Mr. Hempel states that "The response given in Fig. 4 cannot be correct. . .". Perhaps Fig. 4 will be more acceptable in connection with the author's sentence, "We assume that the thermal time constant associated with each slug of water is much longer than the dwell time. . .". If, further, it is remembered that only relative or normalized magnitudes are wanted, then it will be easy to see that Mr. Hempel's final equation under Model 2 reduces to the author's Equation (11). Mr. Hempel's equation incorporates second-order effects which the author purposely omitted in writing his paper.

Mr. Hempel does not see any acceptable degree of accuracy in the analysis of Model 3 which, he admits, is not easy to analyze. He must have forgotten the author's quotation from "How to Hunt a Submarine." The author found gratification in the fact that his simple analysis produced insight into the cause of some of the peculiarities associated with the dynamic behavior of this type of heat exchanger as reported by Masubuchi [8]. For an acceptable degree of accuracy the author refers Mr. Hempel to the painstaking work done by Masubuchi [8].

The second sentence after Equation (27) and the sentence after Equation (29) imply that Model 3 was initially at uniform temperature. Mr. Hempel is right; this was not clearly stated.

Mr. Hempel has overlooked the fact that in a given length of time more heat will transfer from the "very hot" tube slug to the shell-side fluid than can be transferred from the not so very hot shell-side liquid back to the liquid in the tube behind the hot slug. Thus, as a first approximation, to simplify the analysis, the secondary heat transfer has been neglected.

The author used an imaginary impulse, physically not realizable, to inject a great deal of energy into a finite slug of water. The finite but unknown width of the slug of water transformed the initial impulse into a pulse of finite width for the purpose of this problem. Finally, the pulse width was related to C_0. Mr. Hempel recognizes that the point is of minor importance.

The author is glad to agree with Mr. Hempel that Equation (55) for Model 5 can be obtained simply by Fourier-transforming the exponential response and using as the upper limit on the integral infinite time in lieu of time equal to $\tau_d + 4\tau_m$.

The author is in complete agreement with Professor London's

comments and he strongly recommends a careful study of the paper referred to in footnote (8). That very useful paper was brought to his attention shortly after this paper had been submitted for publication.

Professor Takahashi has ably summarized the fundamental concepts upon which the paper is based. The author is grateful to Professor Takahashi for restating the objectives of this paper, and for his kind remarks.

Here the author takes the opportunity to thank Mr. W. I. Caldwell for pointing out, in a private communication dated December 8, 1959, that the dashed phase-angle curve in Fig. 5, marked $-\Omega$, should exhibit discontinuous jumps, from 180 to 360 degrees, again from 540 to 720 deg, and so on. Beyond 180 deg the curve marked $-\Omega$ is therefore not correct in Fig. 5. And Equations (14) and (17) apply only to angles between zero and π radians, 2π and 3π radians, 4π and 5π radians, and so on. As a result the similarity to the Cohen and Johnson theoretical curve is even more striking.

Journal of BASIC ENGINEERING

Information for Authors

• Manuscripts should be submitted in final form to the Editor. Each manuscript must be accompanied by a statement that it has not been published elsewhere nor has it been submitted for publication elsewhere. A paper which would occupy more than 8 pages of the Journal may be returned to the author for abridgment.

• The author should state his business connection, the title of his position, and his mailing address. A short abstract (50 to 100 words) should be included on the first page immediately preceding the introductory paragraph of the paper.

Three copies of the manuscript are required. One of these must be a carefully prepared printer's copy, typed in double spacing, on one side of the page only, with wide margins, on $8^1/_2$ by 11-in. opaque white paper. Mimeographed manuscripts, if prepared with exceptional care, can be accepted provided they are completely edited.

• As far as possible, all mathematical expressions should be typewritten. Greek letters and other symbols not available on the typewriter should be carefully inserted in ink. Care should be taken to distinguish between capital and lowercase letter, between zero (0) and the letter (*O*), between the numeral (1) and the letter (*l*), etc. A letter representing a vector cannot be printed with an arrow above or below it. The letter should be underscored with a single wavy line, wherever it appears in the text, to designate boldface type.

• A list of symbols carefully marked for the use of the editor (thus: κ, Greek l.c. kappa), if it has not been included in the body of the paper, should accompany the manuscript on a separate page.

• Before preparing a manuscript the author should study printed articles in the Journal, with special attention to the form and style of mathematical expressions, tables, footnotes, references, and abstract. Numbers that identify mathematical expressions should be enclosed in parentheses. Numbers that identify references at the end of the paper should be enclosed in brackets. Care should be taken to arrange all tables and mathematical expressions in such a way that they will fit into a single column when set in type. Equations that might extend beyond the width of one column (fractions that should not be broken or long expressions enclosed in parentheses) should be rephrased to go on two or more lines within column width. Fractional powers are preferred to root signs and should always be used in more elaborate formulas. The solidus should be used instead of the horizontal line for fractions wherever possible.

• A normal paper should not exceed 8 pages in the Journal. It is about 8000 words or 20 double-spaced typewritten pages exclusive of drawings. A Technical Brief should not exceed 750 words or about one page in the Journal inclusive of drawings.

• Originals and two copies of figures must accompany the manuscript. Line drawings should not be larger than $8^1/_2 \times 11$ in. and should be planned for reduction to column width. Lettering should be large enough to be clearly legible when the illustration is reduced. The originals of line drawings must be in India ink on white or pale blue tracing paper or tracing cloth. Photographs of equipment or test specimens must be glossy prints and should be used sparingly. Captions for figures should be typed double-spaced and included as the last page of the manuscript. The figure number and author's name should be written in the margin or on the back of each illustration.

• Titles of papers should be brief.

• Authors can obtain copies of the ASME Manual MS-4, "An ASME Paper," from the Editor and are urged to do so before drafting their papers in final form.

• Papers published in Journals of the Transactions of the ASME must be presented at a Meeting, either in person or by title.

• An author is entitled to 25 preprints free of charge (in the case of two authors, 15 each; three or more authors, 10 each). Larger quantities of preprints or reprints can be ordered from Editorial Department, The American Society of Mechanical Engineers, 29 West 39th Street, New York 18, N. Y. Quotations will be sent on request.

Please address all communications to:

Editor, ASME,

29 West 39th Street,

New York 18, N. Y.

ASME Publications Committee